



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Dissertation

Presented to the University of Dublin, Trinity College

in fulfilment of the requirements for the Degree of

Doctor of Philosophy in Computer Science

March 2021

**Applications in Image Aesthetics Using Deep
Learning: Attribute Prediction, Image Captioning
and Score Regression**

Koustav Ghosal

Declaration

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

Koustav Ghosal

August 19, 2021

Permission to Lend and/or Copy

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

Koustav Ghosal

August 19, 2021

Abstract

Image Aesthetics refers to the branch of computer vision which is about the study of aesthetic properties of photographs *i.e.* the factors which make an image look pleasing or dull. Such factors extend beyond the physical properties of an image such as object category or location to subtler and more nuanced ambiguous concepts such as “candid expression”, “harsh lighting”, “bad placement” *etc.* Nevertheless, the problems in Image Aesthetics have traditionally been modelled as classical computer vision tasks such as classification, regression *etc.* And, as with most other problems in computer vision, deep learning based strategies have proved more effective in this area as well, outperforming the classical approaches by a wide margin. Nowadays, automated systems for Image Aesthetics Analysis have widespread applications from professional multimedia content development to casual creatives in social media and advertising.

In this thesis, we study three different applications in Image Aesthetics using deep learning: attribute classification, captioning and score prediction. First, we study the capacity of deep neural networks in capturing the geometric attributes *i.e.* those which depend on the arrangement of objects within the image. Based on this, we propose a system that predicts the dominant aesthetic attributes in a photograph such as The Rule of Thirds, leading lines *etc.* Second, we develop an aesthetic image captioning framework by exploiting *in the wild* user feedback from the web.

Given an image, our framework generates critical feedback such as “*nice composition but the foreground is out of focus*”. Third, we investigate the limitations of traditional convolutional neural networks with respect to global relational reasoning and handling photographs of arbitrary aspect ratio and resolution. We present a visual attention based graph neural network that addresses these limitations and advances the state-of-the-art in aesthetic score prediction.

To my mother and sister.
And father, who would have been proud.

Acknowledgments

It was a privilege to work and learn from my supervisor Prof. Aljosa Smolic every day. This work wouldn't be possible without the perfect balance of freedom and guidance I enjoyed all these years. That he believed in this work, kept it going during the lows. For the most part, I thank him for patiently listening to every unrealistic idea, that I ever had but am now ashamed of.

I thank Gail for taking care of everything else, things probably I am not even aware of so that I could focus on research. Not a single print was taken without dropping in and wasting her time for no reason. She protected and spoiled me, like many others, with that affection.

I am indebted to my past collaborators; Mukta for her insights during the ideation of this work and the motivating long chats; Aakanksha for putting things in perspective when they didn't look great. I would also like to thank other senior members of V-SENSE, who have helped in different ways, time and again; Emin, Iman, Matis, Richard. I thank my transfer reviewers, Prof. Rozenn Dahyot and Prof. Mads Haahr for their valuable feedback. Particularly, I thank everyone in GV2 who participated in the experiments and helped evaluate and strengthen this work. I am grateful to Science Foundation Ireland, Trinity College Dublin and the people of Ireland for providing all the resources that made this research possible.

I thank Prof. Giuseppe Valenzise and Prof. John Dingliana, the reviewers for this thesis, for their time and feedback.

Thanks to my mother and sister, Rina and Rituparna Ghosal, and my brother-in-law Kanan Datta. Growing up and seeing them go out and teach, is probably the foremost reason for why I am in academia; Richik and Sayak, my nephews and stress busters; My cousins Bratati and Sayantan, for being the pillar of support that they are; My friend Tamal, for knowing exactly what to say when I needed to hear it; Parijat for sharing the journey and more; Anupam and Uday Da for introducing me to serious photography and Bresson; Swadesh, Asif, Teesta, Kaustav, Tanushree and Arnab for the lovely memories in Dublin.

And, the lads! Ailbhe, Colm, David, Matt, Pierre, Rafa, Seb, Tejo and Yang. Coming out of the comfort zone to a different culture, I had my share of fear and nervousness about adapting to the new standards. But thanks to these guys, standards were rather very poor!! On a serious note, I have learnt more in the lab while I wasn't working. The Il Capo, the not-so-white-board in *the 1.02*, the shared rejections, the retreats and the nights that I don't remember but very proud of! All of it was a pleasure. Thank you.

KOUSTAV GHOSAL

University of Dublin, Trinity College
March 2021

Contents

Abstract	iv
Acknowledgments	vii
ListofTables	xii
ListofFigures	xv
Chapter 1 Introduction	1
1.1 Motivation	2
1.2 Problems in Image Aesthetics	3
1.3 Research Question and Contributions	7
1.4 Publications	8
1.5 Dissertation Structure	8
Chapter 2 Background	10
2.1 Evolution of Photography	11
2.2 Computational Image Aesthetics	13
2.2.1 Classical Approaches	13
2.2.2 Aesthetic Visual Analysis (AVA) Dataset: <i>Big Data</i> for Image Aesthetics	15
2.2.3 Deep Learning for Image Aesthetics	19
2.2.4 Industrial Applications	21
2.3 Deep Learning for Computer Vision	22
2.3.1 Big Data and Evolution of Deep Learning	22
2.3.2 Network Architectures	24
2.3.3 Vision and Language	27

2.3.4	Learning from Noisy Data	28
2.3.5	Graph Neural Networks	29
2.4	Summary	30
Chapter 3 Geometry Aware Aesthetic Attribute Prediction		31
3.1	Motivation	32
3.2	Network	36
3.2.1	Saliency Detector	36
3.2.2	Feature Extractor	36
3.2.3	Classifier	37
3.2.4	Data Augmentation	37
3.3	Datasets	38
3.4	Experiments	39
3.4.1	Style Classification	40
3.4.2	Per-class Precision Scores	41
3.4.3	Effect of data-augmentation	41
3.4.4	Limitations	41
3.5	Conclusion	50
Chapter 4 Aesthetic Image Captioning from Weakly Labelled Photographs		52
4.1	Motivation	53
4.2	Caption Filtering Strategy	57
4.3	Weakly Supervised CNN	60
4.3.1	Visual and Aesthetic Attributes	60
4.3.2	Latent Dirichlet Allocation (LDA)	62
4.3.3	Relabelling AVA Images	62
4.3.4	Training the CNN	63
4.4	The Final Framework	63
4.5	Experiments	63
4.5.1	Datasets	65
4.5.2	Baselines	65
4.5.3	Results and Analysis	66
4.6	Conclusion	72
Chapter 5 Aspect Ratio and Layout Aware Aesthetic Score Regression		74

5.1	Motivation	75
5.2	Pipeline	79
5.2.1	Feature Graph Construction	79
5.2.2	Score Regression using GNN	81
5.2.3	Training	86
5.3	Experiments	86
5.3.1	Dataset and Metric	86
5.3.2	Ablation study	87
5.3.3	Comparison with the State-of-the-Art	89
5.3.4	Label Imbalance	92
5.3.5	Qualitative Results	93
5.4	Conclusion	95
Chapter 6 Conclusion		97
6.1	Summary	98
6.2	Outlook and Future Work	99
Appendix A Abbreviations		103
Appendix B Aesthetic Attributes		104
Bibliography		108

List of Tables

2.1	Number of samples per category for AVA Style	17
3.1	Style Classification : Comparison with the state-of-the-art : The results are reported in terms of Mean Average Precision (average of per class precision). We observe that for both the datasets, our method performs better than the state-of-the-art . Flickr Style was not used in [1, 2].	40
3.2	PCP for AVA Style Dataset : Sal-RGB outperforms the state-of-the-art [3] by a significant margin in every category. Our own baselines DenseNet [4], ResNet [5], RAPID++ perform equally well for almost all categories except RoT, for which Sal-RGB performs much better.	43
3.3	PCP for Flickr Style dataset : Sal-RGB outperforms the state-of-the-art [3] by a significant margin. Our own baselines DenseNet [4], ResNet [5], RAPID++ perform equally well. The categories in Flickr are mostly appearance based. Hence, no significant improvement is achieved by using Sal-RGB over a regular CNN. Even the <i>geometric composition</i> category contain photographs of objects having regular geometric shapes . Hence, it is not location dependent in true sense.	44
3.4	PCP and MAP for AVA with different augmentation : We observe that a better validation accuracy ensures a better test performance. The decreasing order of mean average precision is as follows : $T_{rnc} > T_w > T_{rc} > T_{icc} > T_{cc}$	46

4.1	(a) Results on AVA-Captions: Both CS and CWS, trained on AVA-Captions perform significantly better than NS, which is trained on nosiy data. Also, the performance of CWS and CS is comparable, which proves the effectiveness of the weakly supervised approach (b) Generalization results on PCCD: Models trained on AVA-C perform well on PCCD validation set, when compared with models trained on PCCD directly. We argue that this impressive generalizability is achieved by training on a larger and diverse dataset.	66
4.2	We measure inter-rater agreement for the scoring strategy using Krippendorff’s alpha. A value between 0 and 1 indicates positive agreement and therefore we find that our strategy is judged by human subjects quite reliably with $\alpha \geq .5$. On the other hand, correlation between the algorithm and human judgement regarding a caption is measured using PLCC and SRCC. We notice that the algorithm proposed is fairly consistent with the human judgement in both the metrics.	71
4.3	Subjective comparison of baselines: We observe that human subjects find CS and CWS to be comparable but both significantly better than NS. This underpins the hypothesis derived from the quantitative results that filtering improves the quality of generated captions and the weakly supervised features are comparable with the ImageNet trained features	72
5.1	Architectural Details	85
5.2	Ablation Study: We start with the most basic single fully connected layer (Avg-Pool-FC) and gradually add the different components namely, the encoder-decoder, feature graph, message passing and readout. We notice steady improvements in the performance in all metrics.	88
5.3	PLCC, SRCC, \mathcal{T}_{Acc}: Our approach outperforms the previous methods for score regression in all the metrics. To the best of our knowledge, [6] is the most recent work on this topic and [7] is the state-of-the-art.	91

5.4 **Accuracy(Acc) and Balanced Accuracy(Acc (B)):** We compare our regression-based approach using indirect thresholded accuracy (\mathcal{T}_{Acc}) with methods which pose the problem as a classification problem and are optimized with a binary classification loss. We find the performance comparable and better in terms of Acc and Acc (B), respectively. 92

List of Figures

1.1	Aesthetic Attribute Prediction: One approach is to predict the probability of a given set of photographic attributes. The values indicate how likely each attribute contributes to the overall aesthetic value of the image. The prediction scores of a classifier are normalized between $[0, 1]$ using a sigmoid function.	4
1.2	Aesthetic Feedback or Captions : The aesthetic captions provide feedback regarding the aesthetic attributes such as lines, light etc. On the other hand the physical captions comment about the physical properties.	5
1.3	Predicted Score / Ground Truth score	6
2.1	(a) <i>Boulevard du Temple</i> , a daguerreotype made by Louis Daguerre in 1838, is generally accepted as the earliest photograph to include people (b) Étienne-Jules Marey’s multiple exposure photography (c) A 360 panorama using a selfie stick (c) Re-focusable light field image captured by Lytro camera (Source : Wikipedia)	11
2.2	Example images from the AVA dataset corresponding to 14 different styles: (L-R) Row 1: Complementary Colors, Duotones, HDR, Image Grain, Light On White, Long Exposure, Macro. Row 2: Motion Blur, Negative Image, Rule of Thirds, Shallow DOF, Silhouettes, Soft Focus, Vanishing Point	16
2.3	User comments from AVA	18
2.4	AlexNet Architecture	24
2.5	VGG-Net Architecture	24

2.6	GoogNet or Inception Architecture	25
2.7	ResNet Architecture	25
2.8	DenseNet Architecture	26
2.9	Standard convolutions vs graph convolutions	29
3.1	Output from our network : A screenshot from our web-based application. Predicted attributes are shown with their probability values (one-vs-all). This is a shot from Majid Majidi’s film ‘The Colours of Paradise’. We see that rule of thirds (for child’s position), shallow depth of field, complementary colours (green background and reddish foreground), image grain (because of the poor video quality) are all well identified.	32
3.2	Our Contributions : (a) Input (col 1), saliency maps (col 2) : Saliency maps are generated using the method proposed in [8]. The position of the main subjects can be obtained from the saliency maps. (b) Our double-column CNN architecture : One column accepts the regular RGB features and the other column accepts saliency maps. The features from RGB channel are computed using a pre-trained Densenet161 [4], fine-tuned on our datasets. They are fused using a fully-connected layer and finally passed to another final fully-connected layer for classification.	34
3.3	Comparison of overall MAP and RoT precision for different networks : We trained ResNet152 [5] and DenseNet161 [4] on AVA Style and Fusion results are from [3]. RAPID++ is implemented following the data augmentation as done in [1] but with Densenet161 architecture. Although the MAP values are not too different, Sal-RGB outperforms others in finding RoT by a significant margin.	42

3.4	Training and validation accuracy (normalized to [0,1]) for AVA Style: For each strategy the proposed two-column network was trained for 30 epochs with a learning rate of 0.001 and a batch-size of 16. We observe, that in terms of overfitting (the gap between training and validation curves), the T_{rnc} and T_w performs best and worst, respectively. The decreasing order of overfitting is observed as follows $T_w > T_{icc} > T_{cc} > T_{rc} > T_{rnc}$. This observation is consistent with [1] where they observe that warping causes overfitting. In our case, both T_w and T_{icc} involve warping and hence are the most overfitted strategies.	45
3.5	Confusion matrix for AVA Style with our model: For a test sample, the rows correspond to the real class and the columns correspond to the predicted class. The values are computed over 2573 test samples of AVA and then normalized. Examples of false positive images can be found in Figure 3.7	48
3.6	Confusion Matrix for our model on Flickr Dataset : Examples of false positive images can be found in Figure 3.7	49
3.7	False positives : Each row corresponds to false positive samples from a pair of mutually confused classes. Column 1-4 and column 5-8 correspond to the first and second category in a pair, respectively. Top-Bottom - Long Exposure / Motion Blur, Shallow DOF / Macro, Shallow DOF / Bokeh, Geometric Composition / Minimal, Horror / Noir, Pastel / Vintage	50
4.1	Aesthetic image captions. We show candidates generated by three different frameworks discussed in this chapter: (a) For NS, we use an ImageNet trained CNN and LSTM trained on noisy comments (b) For CS, we use an ImageNet trained CNN and LSTM trained on compiled AVA-Captions dataset (c) For CWS, we use a weakly-supervised CNN and LSTM trained on AVA-Captions	54
4.2	Informativeness of captions.	57
4.3	Some topics/labels discovered from AVA-Captions using LDA.	61
4.4	Proposed pipeline	64

4.5	Diversity: Figures (a) - (c) report diversity of captions following [9]. The x -axes correspond to n -gram positions in a sentence. The y -axes correspond to the number of unique n -grams at each position, for the entire validation set. Figure (d) plots the overall diversity, as reported in [10]. We observe that the diversity of the captions increase significantly when the framework is trained on cleaner ground-truth <i>i.e.</i> AVA-Captions (CS or CWS) instead of AVA-Original (NS).	68
4.6	Subjective evaluation of caption filtering: The matrix compares our scoring strategy and human judgement for distinguishing a <i>good</i> and a <i>bad</i> caption. The rows stand for our output, and the columns represent what humans thought. We observe that the proposed caption filtering strategy is fairly consistent with what humans think about the informativeness of a caption.	70
5.1	The proposed two stage pipeline	75
5.2	Message Passing with Self-Attention: A toy scenario for the update $v_i \rightarrow v'_i$, with four neighbours and three attention heads (red, blue and black). v'_i is the concatenation of the output from the different attention heads (Eq 5.6). Note, that this step is repeated for every node $v_i \in G$ and the output is also a graph with the same structure as the input.	82
5.3	(a)-(f) Average score distribution of different baselines plotted with the ground truth distribution. (g)-(h) Baseline predictions on five random images from the test set	90
5.4	Confusion Matrices	94
5.5	GAT_{×3}-GATP predictions / Ground truth (GT) scores for images randomly sampled from AVA:	96
6.1	Correlated Attributes: Each row displays eight samples from two categories (four each) which were mutually confused. Row 1 is horror / noir and Row 2 is minimal / geometry	99
6.2	Realistic Comments	100
B.1	Colour Wheel : Example of complementary colours are red-green, yellow-purple combinations.	104

B.2 Rule of Thirds : It is observed that when the main subject is placed at one of the four points instead of the centre of the photograph, it is aesthetically more pleasing. 106

B.3 Some photographic attributes from AVA and their descriptions107

Chapter 1

Introduction

In this chapter, we aim to introduce the reader to the field of Image Aesthetics and the motivation for studying the problem. A detailed discussion on the general principles of aesthetics in photography from an artistic perspective is beyond the scope of this work. Instead, we follow a top down approach and try to highlight the aspects relevant for discussing the technicalities of the applications explored. We discuss the challenges involved, key research questions explored, primary contributions and a high-level structure of the thesis.

1.1 Motivation

What makes a good photograph?

This is a fundamental question that intrigues anyone into serious photography. It is quite common for photographers to take a picture with an idea but being unable to convey that to the viewer. The rules of photographic composition are important in this context. Experienced photographers are generally aware of those rules. They apply, combine and extend these rules to guide the viewer to the subject of the photograph. One starts by learning simple concepts such as keeping the horizon straight, focusing the subject sharply, getting the correct exposure *etc.* and then with practice, grasps the more complex rules such as The Rule of Thirds, complimentary colours, vanishing lines *etc.* [11]¹. The rules of composition are important not only to the photographers but also to the viewers in order to appreciate a photograph.

Can an artificial agent be trained to criticize a photograph?

How difficult could this task, which is reasonably challenging for humans, be for a computer? The computer vision community has tried to imitate the human visual system for certain tasks such as recognition [12, 13], detection [14, 15], segmentation [16] *etc.* Moreover, since the last decade Deep Learning [17] has been quite successful in solving these classical problems. The data-driven deep features are more robust than the hand-crafted features such as HOG [18], SIFT [19] *etc.* in terms of modelling the physical properties such as class, colour, shape, position *etc.* But how well do these approaches map to the more subtle and complex aesthetic properties of a photographic image? Can a computer be made to understand the sharpness of focus or the harshness of light? Can we train an artificial agent to provide feedback to amateur photographers? Or can a computer rate or score a photograph on a given scale? In other words, as put by Aaron Hertzmann [20], can computers understand or create art (photography)? In this thesis, we explore these possibilities.

¹Please check Appendix B for more on these rules

Big Data and Deep Learning for Image Aesthetics

As just mentioned, in recent years data-driven solutions have taken over most problems in computer vision and this applies to Image Aesthetics as well. In the past, non deep learning or feature-based attempts tried to identify and encode the aesthetic properties and define a generic model for Image Aesthetics [21, 22, 23, 24, 25, 26]. In most cases, these features were extensions of the hand-coded off-the-shelf generic image descriptors such as HOG, SIFT *etc.* But, because of the ambiguous and overlapping nature of aesthetic properties, the task was quite complex and ill-posed.

In recent years, deep learning based data-driven approaches have proven quite effective [7, 1, 2, 27, 28]. This is primarily due to the availability of the large scale Aesthetic Visual Analysis (AVA) [29] dataset and the rapid improvements in convolutional neural network (CNN) architectures. It is not at all surprising that the current state-of-the-art methods in Image Aesthetics outperform the classical approaches by a wide margin. Nevertheless depending on the task, applying deep learning for Image Aesthetics is not straightforward and has its limitations. In the next section, we introduce the applications studied in this thesis and the associated problems.

1.2 Problems in Image Aesthetics

Specifically, we look at three different applications in Image Aesthetics: Attribute Classification, Feedback or Captioning and Score Prediction.

Attribute Classification

As discussed in Sec. 1.1, a photographer considers different factors or attributes to judge the aesthetic value of an image. These attributes could be appearance-based such as exposure, texture, colour balance *etc.* or geometry/layout based such as arrangement of subjects, negative space, The Rule of Thirds *etc.* A combination of these factors applied correctly makes a composition look aesthetically pleasing or dull. However, it is worth mentioning here that no list of such factors can be exhaustive. Eval-

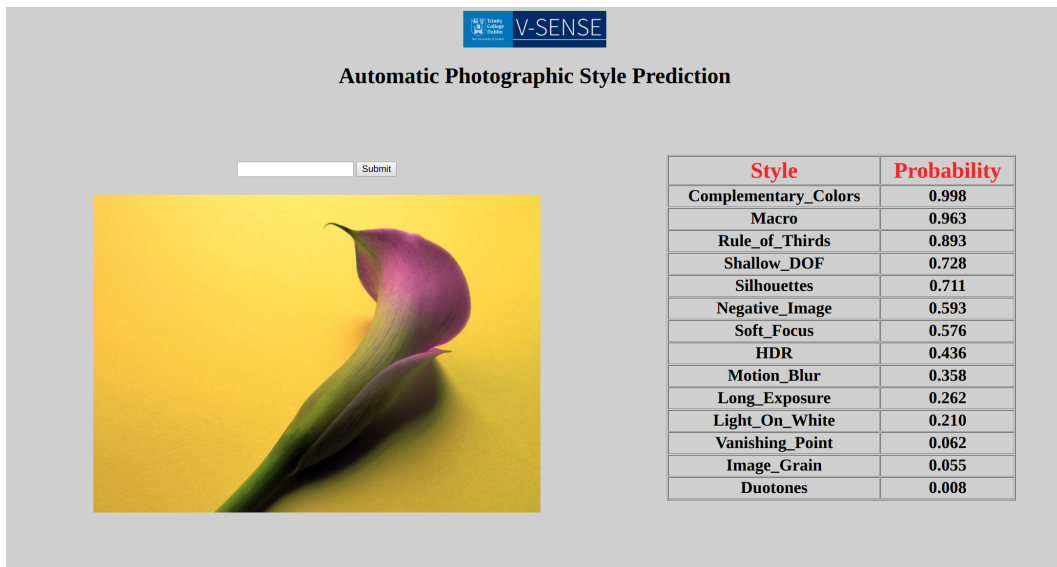


Figure 1.1: **Aesthetic Attribute Prediction:** One approach is to predict the probability of a given set of photographic attributes. The values indicate how likely each attribute contributes to the overall aesthetic value of the image. The prediction scores of a classifier are normalized between $[0, 1]$ using a sigmoid function.

uating a photograph is a complex and subjective task. As in many other artistic disciplines, contradictory opinions regarding the aesthetic attributes of a photograph is quite common among critics. This makes the problem severely ill posed. Instead, we address a simplified version of the problem and study the contribution of N different aesthetic attribute as shown in Fig. 1.1.

Based on the recent developments of CNNs, we develop a system that predicts the probability distribution over the set of N aesthetic attributes. However, CNNs by design are translation invariant *i.e.* they are limited in their ability to capture object placement. In other words, they struggle to understand the significance of the placement of subjects within a photograph. But such an understanding is crucial for certain "geometric" attributes such as The Rule of Thirds, which states that placing the subjects in a certain way makes a photograph more appealing. In Chapter 3, we propose a visual saliency based double column framework that addresses this issue and achieves the state-of-the-art results in aesthetic attribute

		
i like the composition and the lines of the building	i like the way you have captured the light	i like the way the sun is shining through the clouds
barricade in front of stairs	a multi-storeyed building	a line of trees during sunset

Figure 1.2: **Aesthetic Feedback or Captions :** The **aesthetic captions** provide feedback regarding the aesthetic attributes such as lines, light etc. On the other hand the **physical captions** comment about the physical properties.

classification.

Feedback/ Captioning

By feedback, we mean a description of the photograph in an *informative* and intelligible format. For example, in photography websites such as Flickr or Dpchallenge, experts provide critical feedback to photographers based on their evaluation, such as “*I like how the lines of the road lead the eyes to the main subject*”. We emphasize on “informative” to filter out the comments made by non-experts which are less useful for the photographer such as “*nice shot*” or “*well done*”. Unless there is a detailed explanation of what makes the picture aesthetically pleasing or unpleasant, the feedback is less useful [30]. Therefore as the target task, an agent should be able to generate informative aesthetic descriptions of the photograph.

In Chapter 4, we present our study about Aesthetic Image Captioning. Motivated by the recent developments in traditional image captioning *i.e.* textual descriptions of the physical properties of natural scenes (*E.g.* “*A red car in front of a building*”), we develop a pipeline based on a CNN feature extractor and a long and short term memory (LSTM) network. However,

the lack of a curated dataset makes the task quite challenging and we exploit noisy “*in the wild*” comments from the web to build the framework. Furthermore, we compile a dataset called AVA-Captions which consists of about 230,000 images with 5 captions each, on an average.

Score Prediction

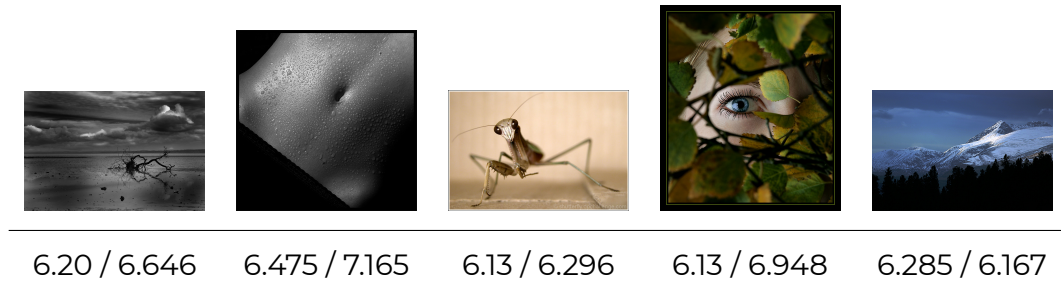


Figure 1.3: Predicted Score / Ground Truth score

Predicting a rating or score of an image on a given scale is the most researched problem in Image Aesthetics. Unlike the attribute classification problem which aims to isolate the dominant aesthetic elements in an image, a score is an *overall* measure of its aesthetic appeal. In the context of casual creatives, it is analogous to the popularity of an image on social media such as the number of likes on Facebook or Instagram. Industrial applications of score prediction include ranking and sorting images for marketing and advertisement, smarter cameras and editing software etc.

However, most deep learning-based methods for aesthetic score prediction face two primary challenges — aspect ratio awareness and image layout understanding. The aspect ratio of the photographs gets distorted while they are resized/cropped to a fixed dimension to facilitate training batch sampling. On the other hand, the convolutional filters process information locally and are limited in their ability to model the global spatial layout of a photograph. In Chapter 5, we present a two-stage framework based on graph neural networks and address both these problems jointly. First, we propose a feature-graph representation in which the input image is modelled as a graph, maintaining its original aspect ratio and resolution. Second, we propose a graph neural network architecture that takes this

feature-graph and captures the semantic relationship between different regions of the input image using visual attention.

1.3 Research Question and Contributions

In this section, we formally present the research question studied in this dissertation and summarize the key contributions discussed above.

Research Question

Broadly, we investigate — “***How efficiently can artificial agents be trained for Image Aesthetic Analysis?***”.

We study this problem in the context of three objectives:

- Aesthetic Attribute Prediction.
- Feedback or Aesthetic Image Captioning.
- Aesthetic Score Prediction.

Contributions

- We study the capacity of convolutional neural networks for understanding the geometric aesthetic attributes and develop a visual saliency based framework for **attribute classification**. The proposed method achieves state-of-the-art results on several datasets.
- We present a framework for **aesthetic image captioning** using the *weakly* labelled data from the web. We build our framework based on a traditional CNN-LSTM pipeline used for traditional image captioning. We compile a new benchmark dataset for aesthetic image captioning called **AVA-Captions**.
- We propose a graph neural network based aesthetic score predictor which is aspect ratio and layout aware. Our method advances the state-of-the-art in **aesthetic score regression** on the Aesthetic Visual Analysis (AVA) benchmark dataset.

1.4 Publications

Publications Based on Thesis Work

- Koustav Ghosal, Mukta Prasad, Aljosa Smolic, **A Geometry-Sensitive Approach for Photographic Style Classification**, Irish Machine Vision and Image Processing Conference, August 2018 (IMVIP), Belfast.[31]
- Koustav Ghosal, Aakanksha Rana, Aljosa Smolic **Aesthetic Image Captioning From Weakly-Labelled Photographs** The IEEE International Conference on Computer Vision (ICCV-W), 2019 [32]
- Koustav Ghosal, Aljosa Smolic, **Aspect Ratio and Spatial Layout Aware Image Aesthetics Assessment Using Graph Attention Network**, IEEE Transactions on Image Processing (under review)

Publications Outside the Scope of the Thesis

- Xu Zheng, Tejo Chalasani, Koustav Ghosal, Sebastian Lutz, Aljosa Smolic **STaDA: Style Transfer as Data Augmentation** 14th International Conference on Computer Vision Theory and Applications, 2019. [33]
- Ojasvi Yadav, Koustav Ghosal, Sebastian Lutz, Aljosa Smolic, **Frequency Domain Loss Function for Deep Exposure Correction of Dark Images**, Signal Image and Video Processing [34]

1.5 Dissertation Structure

This dissertation is divided into six chapters and two appendices. In this first chapter we discussed the motivation and problems addressed in the thesis. In Chapter 2, we discuss related work. We provide a detailed review of classical and recent work on Image Aesthetics. We also briefly mention

research which may not be directly related but has influenced certain aspects of our work. In the next three chapters we discuss the three applications we explored. In Chapter 3, we present our work on aesthetic attribute classification, in Chapter 4 we discuss aesthetic image captioning and in Chapter 5, we talk about our work in aesthetic score prediction. Each of these chapters stand on their own and can be read independently. The ordering is based on a rough chronological order followed during this thesis. In Chapter 6, we conclude this thesis by summarizing the key contributions and future research directions.

Chapter 2

Background

In this chapter, we discuss previous research in Image Aesthetics and other related areas. Owing to the diverse applications explored, the scope of this dissertation is quite broad and spans multiple modalities (image and text). There is a plethora of literature in almost every topic covered. We try to be exhaustive and inclusive in reviewing the classical and recent developments in Image Aesthetics. For other related areas, we restrict the discussion to the papers that are recent and most relevant to the problems addressed. We refer to several survey papers during the discussion which may be of interest for further exploration.

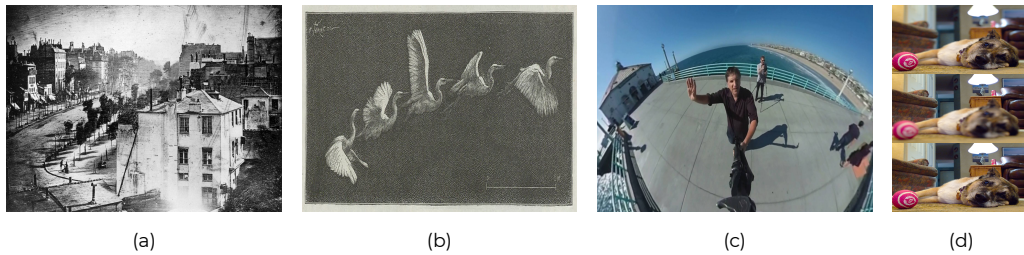


Figure 2.1: **(a)** *Boulevard du Temple*, a daguerreotype made by Louis Daguerre in 1838, is generally accepted as the earliest photograph to include people **(b)** Étienne-Jules Marey’s multiple exposure photography **(c)** A 360 panorama using a selfie stick **(d)** Re-focusable light field image captured by Lytro camera (Source : Wikipedia)

2.1 Evolution of Photography

Photography: Art or Not?

The acceptance of photography as an art form did not come easily. In its early years, it was used merely as a tool to preserve practical records of the world by using techniques such as Louis-Jacques-Mandé Daguerre’s *daguerreotype* (Figure 2.1(a)) or William Henry Fox Talbot’s *negative-positive process* [35, 36]. But as technology improved over the years, it started being seen as an alternative to contemporary realist paintings. The key parameter of criticizing a realist painting was its resemblance to reality *i.e.* how well it mimicked the real world. With cameras doing that job reasonably well, painting started evolving towards newer styles. Photographic techniques such as Étienne-Jules Marey’s multiple exposure (Figure 2.1(b)) had a profound impact on the then avant-garde genres of Futurist and Cubist paintings. With the borders between painting and photography being crossed from either side, the debate, whether photography should be recognized as an art form, was inevitable. The formal recognition was achieved in 1910 when the first photographic exhibition “Buffalo Show” by Alfred Stieglitz was organized at The American Art Museum [20, 37].

Missing the *Human Element*

One of the arguments against photography was the mechanical process behind its creation. In the beginning, photography was viewed primarily

as a photo-chemical process used for record-keeping rather than an artistic practice. While there are many different definitions of art, the common factor is the existence of an artist who creates the artefact. Islamic forms and geometric patterns from the past are considered art works but the natural patterns in clouds and rocks are not. It is the presence of the human element that makes the distinction [38]. Similarly, it is the use of machines to do most of the job that challenged photography's artistic status. Nonetheless, it aroused great interest and photographers pushed the boundaries by introducing new processing techniques and genres [39]. Gradually, photography earned the status of a traditional art form. Today it is pursued not just as a hobby, but as a powerful medium for creating social awareness regarding human-rights violations, wildlife preservation, dangers of war, etc. [40]. At the same time, the new inventions in camera technology such as light-field, panoramic, 360 cameras, etc. [41, 42] (Figure 2.1 (b) and (c)) are opening new possibilities.

Photography Criticism and Image Aesthetics

One of the main reasons for its rising popularity is the easy access to a decent camera. With little or no formal training in photography and the help of an automatic or semi-automatic camera, it is possible to capture *accidental* pictures which are aesthetically quite pleasing. Nowadays, almost all the good cameras and even mobile phones come with advanced hardware and software that allows the photographer to capture pictures under harsh conditions such as low light, motion etc. But, for someone new to photography, there is a need to be criticized and evaluated in order to make progress [30]. Apart from assisting photography enthusiasts, criticism plays an important role in advertising and creative industries for content creation and management.

Research in Image Aesthetics is primarily focused at automating this process. An *intelligent* assistive technology for the camera or smart softwares for organizing and filtering creative content are some of the many real-world applications which can be potential spin-offs from this research. In the next section, we discuss the classical and recent developments in this area along with some of its industrial applications.

2.2 Computational Image Aesthetics

We divide this section into four parts. In Section 2.2.1, we talk about the classical approaches, in Section 2.2.2 we introduce the AVA benchmark, in Section 2.2.3 we discuss the deep learning based methods and finally in Section 2.2.4 we discuss some of the recent industrial applications.

2.2.1 Classical Approaches

Early work in Image Aesthetics relied on explicit-modelling or hand-coding of popular aesthetic attributes. Many of these approaches tried to jointly capture global properties such as overall colour harmony, illumination and contrast and local factors such as contrast between objects, blur and clutter. [21, 22, 23, 24, 25, 26, 43, 44, 3, 45, 46, 47, 48]. In this section, we discuss each of these briefly and in the end, point to a survey for further exploration.

Datta *et al.* [21] in their pioneering work on Image Aesthetics use images from Photo.net and hand-code 56D features for modelling nine different aesthetic properties — exposure, saturation, The Rule of Thirds, familiarity, size and aspect ratio, region composition, depth of field and shape convexity. They train a support vector machine for aesthetic category classification and score regression.

Ke *et al.* [22] identify three sets of high-level perceptual factors which distinguish high-quality professional photographers from the low-quality snapshots. The factors are simplicity (lack of clutter and salient foreground), realism (mundane content vs surreal processing) and basic techniques (blur, contrast *etc.*). They encode these factors using low level features such as edge, colour, hue, intensity distributions and frequency components and train a Naive Bayes classifier.

In [23], Luo *et al.* present a method in which the main subject in an image or video is focused for feature extraction rather than the entire image. For images, they define features pertaining to blur, clarity contrast, simplicity of absence of clutter, composition geometry and colour harmony. For videos, they add subject motion and overall motion stability. Using these features they train Naive Bayes, SVM and Adaboost classifiers and achieve

state-of-the-art results for the task.

A similar approach was adopted by Obrador *et al.* in [24] where they propose three types of features: *simplicity* features based on image segmentation, *global* features or an overall measure of luminance, contrast, colourfulness, harmony and composition geometry and *low-level* features based on contrasting regions in the image. They create a dataset of 2100 images with photographs from the Dpchallenge website having seven categories and train classifiers using SVM.

Luo *et al.* [43] propose a genre specific approach in which they manually divide photographs into different categories such as animal, plant, architecture *etc.* and then extract regional and global features separately for each category. The idea was motivated by the fact that human beings treat each genre differently while analyzing the aesthetic appeal. In addition they also compile the CUHK dataset of 17613 images.

Motivated from how humans analyze a photograph, Dhar *et al.* [25] propose three types of high-level describable features based on composition (image layout and geometric attributes), content (object and scene category) and sky illumination (overall measure of illumination). Using these features, they present an approach for measuring the aesthetic quality and interestingness of photographs.

Marchesotti *et al.* [44] suggest an alternative approach to aesthetic assessment using 'generic' bag-of-visual-words and Fisher Vector based features used in traditional image classification tasks at the time. Their approach was motivated by the computational cost and lack of generalizability of the previous hand-crafted features.

Karayev *et al.* [3] in their experimental work compare several such generic image descriptors for aesthetic attribute prediction. Specifically, they compare LAB colour histogram, GIST, graph-based visual saliency, meta class binary features and deep features on three different datasets; AVA, Flickr Style and Wikipaintings.

Aydin *et al.* [45] propose a system which predicts the contribution of some photographic attributes towards the overall aesthetic quality of a picture. After estimating the extent of certain compositional attributes, they ag-

gregate the scores for different attributes to predict the overall aesthetic score of a photograph by using a novel calibration technique.

Li *et al.* [46] propose a method to assess the aesthetic value of digital copies of paintings based on a user study where they ask artists to list the factors driving a composition such as colour, composition, meaning/content, texture/brushstrokes *etc.* They design global and local features exploiting colour distribution, brightness, blur, shape of segments, contrast between segments *etc.* and achieve results comparable to human performance.

A specific application of aesthetic assessment for portrait shots is presented by Li *et al.* in [47]. They develop a system with three capabilities: aesthetic assessment based on colour, lighting, composition and facial characteristics, a cropping based photo editing algorithm and a retrieval framework from a large collection of consumer photos.

In [48], Yao *et al.* present OSCAR, a comprehensive assistive software for photography enthusiasts. Their system helps users in three different ways. It suggests exemplar images from a database in terms of content and composition, provides a confidence score regarding the colour correctness and finally, an overall aesthetic score.

Survey

In their 2011 tutorial [26], Joshi *et al.* present a comprehensive survey of the research and challenges in analyzing the aesthetics and emotions of photographs. They outline the key philosophical and artistic motivations behind this research, list datasets, survey contemporary research and point to future research problems.

2.2.2 Aesthetic Visual Analysis (AVA) Dataset: *Big Data* for Image Aesthetics

One of the challenges for these early methods was the lack of a standard benchmark for evaluation. Typically, a small set of images were sampled from photography websites such as Dpchallenge.com or Photo.net and used for the specific task at hand. In 2012, Murray *et al.* in their semi-

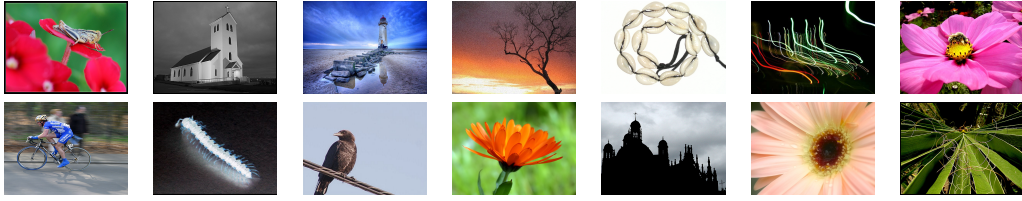


Figure 2.2: **Example images from the AVA dataset corresponding to 14 different styles:** (L-R) **Row 1:** Complementary Colors, Duotones, HDR, Image Grain, Light On White, Long Exposure, Macro. **Row 2:** Motion Blur, Negative Image, Rule of Thirds, Shallow DOF, Silhouettes, Soft Focus, Vanishing Point

nal work ‘AVA: A Large-Scale Database for Aesthetic Visual Analysis’ [29] compiled the first large-scale publicly available benchmark for the task. This was also the same year in which AlexNet [12] won the Imagenet Large Scale Visual Recognition Challenge (ILSVRC) [49] and the deep learning wave started to rise. The release of AVA and the subsequent rapid evolution of deep learning together led to a significant volume of work in Image Aesthetics over the years. We discuss those in detail in the next section. But first, in this section we discuss the details of the AVA dataset.

AVA is a collection of ~255,000 images crawled from www.dpchallenge.com. Dpchallenge is a curated forum for photographers where challenges are hosted based on certain themes such as ‘bokeh’, ‘complimentary colours’, ‘leading lines’ etc. (Figure 2.2). Participants post pictures and the wider community rates each photograph during the challenge on a scale of 10 and posts comments during and after the challenge. There are several types of ‘weak’ annotations available with AVA, although some of them are noisy and non-curated.

Ratings

The user ratings for the images are provided as a 10-bin histogram of scores. The final score is obtained as a weighted average of the histogram. There are two traditional approaches for aesthetic analysis using AVA. One is to solve a classification problem by thresholding the scores at 5 and generate ‘good’ and ‘bad’ labels. The other approach is to regress the scores directly. However, a fundamental problem with these ratings is the uneven

Index	Classes	Number of Samples			
		Train	Val	Test	Total
1	Complementary_Colors	622	138		
2	Duotones	843	198		
3	HDR	277	40		
4	Image_Grain	577	95		
5	Light_On_White	794	166		
6	Long_Exposure	563	113		
7	Macro	1138	221		
8	Motion_Blur	415	73		
9	Negative_Image	641	127		
10	Rule_of_Third	923	189		
11	Shallow_DOF	994	190		
12	Silhouettes	694	130		
13	Soft_Focus	468	100		
14	Vanishing_Point	442	98		
Total		11269	2516	13785	

Table 2.1: Number of samples per category for AVA Style

distribution of scores. Since participants post their best pictures during a competition, the dataset is biased towards ‘good’ labels with a 7/3 ratio. As a result an ‘all positive’ classifier trained on AVA could easily achieve 70% overall accuracy. We talk more about this and possible solutions in Chapter 5. The classification/regression problem based on these ratings is currently the most widely researched topic in Image Aesthetics.

Tags

Of the 250,000 photographs in AVA, the authors manually select 72 challenges, corresponding to 14 different photographic styles as illustrated in Figure 2.2 and create a subset called AVA Style containing about 11,000/2500 training/test images, respectively. The ground truth annotations are provided as one-hot encoding of the style tag associated with an image. Notably, the training data is single labelled whereas the test data is multi-labelled. As with the rating distribution, the tags are not evenly distributed and therefore the number of samples in certain categories is quite sparse (Table 2.1). Aesthetic attribute prediction is a popularly addressed task using the AVA-Style dataset. All the 255,000 images are associated with a tag


Image	Comments
	Photo Quality : Awesome
	I love the colors here
	I like the trees in the background and the softness of the water.
	The post processing looks great with the water, but the top half of the photo doesn't work as well.
	Its a lovely scene but the PP is just a bit over the top for me, the sky pulls me away from the green, a slightly less enhanced shot would have been fine for me. It looks like the lake district or somewhere like that.

Figure 2.3: User comments from AVA

but not all of them are a part of the AVA Style dataset.

Comments

Typically, in Dpchallenge, users ranging from casual hobbyists to expert photographers provide feedback to the images submitted and describe the factors that make a photograph aesthetically pleasing or dull. In a way, they are similar to the captions available in the well instructed and curated datasets such as MC-COCO [50]. But the raw comments in AVA are unconstrained user-comments in the wild with typos, grammatically inconsistent statements, and also containing a large number of comments occurring frequently without useful information (Figure 2.3). Nevertheless, they contain rich information, which has often been used as auxiliary data to boost the performance of the primary tasks. In Chapter 4, we discuss more on the comments and present a framework that exploits these comments for aesthetic image captioning.

To summarize this section, AVA is the largest benchmark available to train and evaluate frameworks for Image Aesthetics but it is weakly labelled. In other words, the ground-truth annotations available for AVA were not

curated and hence unconstrained. As a result AVA closely represents real world data with all the standard challenges such as noise, label imbalance etc. However, it has been subsequently used for diverse image aesthetic assessment tasks by most recent works, if not all of them.

2.2.3 Deep Learning for Image Aesthetics

As in many other computer vision problems, deep learning based methods have been extensively explored in the domain of image aesthetic assessment as well. With the availability of AVA and rapid advances in neural network architectures, the current state-of-the-art methods outperform the classical approaches by a wide margin [1, 2, 51, 28, 52, 27, 53, 54, 55, 56, 57, 7, 6, 58].

In [1], Lu *et al.* propose a double column CNN architecture, where the first column accepts a cropped patch from the input image and the second column accepts a warped version of the entire input. Their intuition was that both local and global information from an image are crucial for aesthetic assessment. In subsequent work [2], multiple patches are cropped from an input and forwarded through the network. The features from multiple patches are aggregated before the final fully-connected layer for classification. The authors argue that sending multiple patches from the same image encodes more global context than a single random crop.

Kong *et al.* [51], propose a ranking framework by training a CNN using Siamese Triplet Loss instead of categorizing images to a coarse binary label. Additionally, they compile the Aesthetics and Attribute Database (AADB) which consists of 10000 images and ground-truth annotations, similar to AVA.

Mai *et al.* in their work [28] propose a network that uses a composition-preserving input mechanism. They introduce an aspect-ratio aware adaptive pooling strategy that reshapes each image differently and thereby claims to preserve the aspect-ratio information, a key element in photographic composition.

In [52], the authors propose a network that predicts the overall aesthetic score and eight style attributes, jointly. Additionally, they use gradient-

based feature visualization techniques to understand the correlation of different attributes with image locations.

Ma *et al.* [27] follow a multiple-patch extraction approach, and the patches are selectively extracted based on saliency, pattern-diversity and overlap between the subjects. Essentially, these techniques attempt to incorporate global context into the features during a forward pass either by warping the whole input and sending it through an additional column or by providing multiple patches from the input at the same time.

Hii *et al.* [54] exploit the textual comments in AVA and augment it with the visual features. The visual features are extracted by appending global average pooling layers on top of multiple Inception blocks and the text is encoded using a recurrent neural network. They also present visualizations of the learnt network activations.

Sheng *et al.* [55] propose a multiple patch aggregation strategy for binary label classification. Using three different forms of visual attention, their network learns to select relevant patches from the input.

Talebi *et al.* [56] propose a new loss function based on the Earth Mover Distance and apply it for aesthetic score regression. Using that, they train several backbone architectures such as Inception, MobileNet *etc.* and apply them for aesthetic as well as no-reference Perceptual Quality Assessment.

Liu *et al.* [57] explore dilated convolutions and graph neural networks for encoding global properties in photographic composition. Features are extracted using a pre-trained DenseNet backbone and then are fed through a few dilated convolution layers before passing to a graph layer for capturing long range dependencies. They also propose a novel loss function which handles the label imbalance in AVA for binary classification.

Hosu *et al.* [7] propose multi-level spatially pooled features for aesthetic score regression. They have a two stage framework. First, features are extracted using an Inception-Resnet backbone and saved to the disk. Then these features are trained for score regression using an inception layer followed by 3 fully connected layers.

Xu *et al.* [6] use spatial attention to focus on certain areas of the image that contribute towards the overall aesthetic value of an image. Their self-attention module is configured on top of an Inception backbone to process features selectively and finally a classifier is trained using Earth Mover Loss for score regression.

Chen *et al.* [58] propose an adaptive fractional dilated convolutional network capable of handling inputs of arbitrary aspect ratios. In order to achieve efficient mini-batching for samples of arbitrary sizes, they present a grouping strategy that reduces computational overhead, significantly.

Survey

Image Aesthetic Assessment: An Experimental Survey [53] by Deng *et al.* is an extensive survey of Image Aesthetics in recent years. It lists many important papers until 2017, proposes a balanced sampling strategy for handling label imbalance in AVA and provides several key insights in this area.

2.2.4 Industrial Applications

Data-driven technologies are being applied for developing or enhancing diverse commercial applications around Image Aesthetics. For example, professional editing softwares such as Adobe Photoshop [59] or Snapseed [60] have automatic filters for aesthetic quality improvement. Several new features based on machine learning are regularly being added to these tools such as subject/object selection, content-aware filling, sky-replacement *etc.* [61]. Understanding what works best for a given photograph is central to automating/enhancing these tools. Casual creative applications in popular domains such as Instagram [62] or Facebook [63] can also improve using this technology for similar reasons.

Another sector which has an ample scope for automated Image Aesthetics is *Stock Photography*. Stock photography agencies act as mediators between photographers and businesses seeking photographs for advertising or marketing. Several websites such as Shutterstock [64], EyeEm [65], Everypixel [66] automatically rank and tag the submitted photographs catering to the need of their customers. Everypixel even has a neural net-

work based automated score estimator where users can verify the aesthetic appeal of their photos.

2.3 Deep Learning for Computer Vision

In this section, we review the topics which are related to our central theme *i.e.* deep learning based Image Aesthetics. All these topics are quite active areas themselves and have a plethora of publications. In the following sections we discuss papers that directly and indirectly motivated our research. In Section 2.3.1, we briefly highlight the evolution of deep learning over the years. In Section 2.3.2, we briefly talk about the neural network architectures which have been used as backbone frameworks in the subsequent chapters. In Section 2.3.3, we discuss multimodal problems in vision and language and in Section 2.3.4 we discuss deep learning applications with nosiy real-world data, both of which have influenced our work in aesthetic image captioning in Chapter 4. In Section 2.3.5, we explore the recent advances in graph neural networks which is the basis for our contributions in Chapter 5.

2.3.1 Big Data and Evolution of Deep Learning

Most problems in Computer Vision (or machine learning in general) can be formulated as function estimation tasks. Given a task and well defined source and target domains, the challenge is to learn an efficient mapping from the source to the target (such as image to labels or image to segmentation masks). Early function estimators such as support vector machines [67] or Naive Bayes [68] were typically optimized based on hand-coded image descriptors such as SIFT [19] or GIST [69]. But these features were based on intuition and assumptions regarding data and thus limited in their ability to capture the diversity of *in the wild* real world image distributions. In other words, these features lacked generalizability and usually performed well for small curated datasets.

Neural networks started becoming popular as they make no strong assumptions about the data distribution and can be optimized using the raw data and labels in an end-to-end fashion using backpropagation. Early

neural networks such as multi layered perceptrons (MLP) showed promising results on image classification problems [70]. But, the number of parameters in MLPs grew exponentially with the input size and hence they were not suitable to handle high dimensional RGB data. But with the rise in popularity of social media and smartphones image data became high resolution, cheap and abundant. This called for a computationally efficient and scalable version of neural networks which could be applied to image data to develop real world applications.

CNNs were able to handle this efficiently using filters with shared weights. A typical filter in a CNN is shared across the entire input and thus the number of parameters is independent of the input dimension. Additionally, it has several other benefits such as translation invariance, hierarchical representation etc. [71]. Also mathematically, the working principle of a CNN could be reduced to a series of matrix operations resulting in fast and efficient GPU implementations [72, 73]. Over the years, as data became more abundant, CNNs grew deeper and GPU technology too improved rapidly to cater for the deep learning community. The result is the complete paradigm shift in artificial intelligence research from model-based to data-driven approaches. Problems such as MNIST handwritten digit classification [70], which were considered quite challenging earlier were solved and became saturated. Many efficient real world applications in areas such as face and gesture recognition, self driving cars, biomedical imaging etc. were developed. In the next section, we briefly review some of the popular CNN architectures which have motivated this work.

2.3.2 Network Architectures

AlexNet

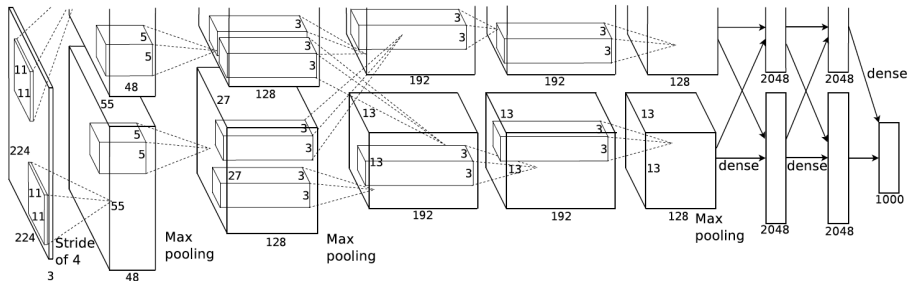


Figure 2.4: AlexNet Architecture

In 2012, Krizevsky *et al.* [12] proposed AlexNet (Figure 2.4), which won the ILSVRC [49] challenge by defeating the first runner up by a significant error margin of 10.8%. The authors showed that depth was a crucial factor that affected the efficiency of CNNs. It consists of eight layers - five convolutional layers followed by three fully connected layers. They also introduced max-pooling and ReLU activations, two of the key concepts widely used in CNNs.

VGG-Net

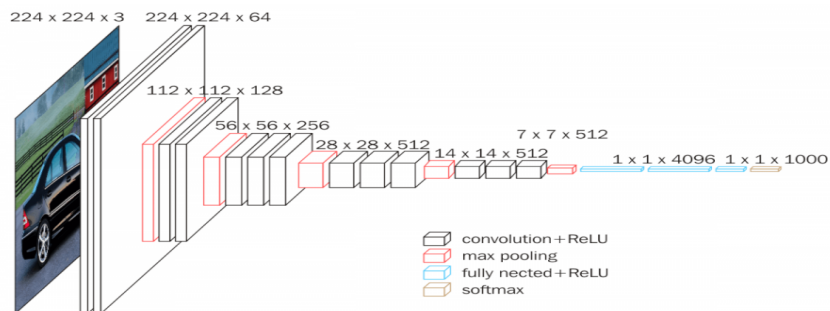


Figure 2.5: VGG-Net Architecture

While AlexNet used large receptive fields (11×11 filters in the first layer), Simonyan *et al.* [13] focused on depth and proposed the VGG-Net (Figure 2.5). VGG-Net uses smaller receptive fields (3×3) and more layers. Using a smaller receptive field made it computationally more efficient and adding

more layers made it learn hierarchical representations. It came as the first runner up in ILSVRC 2014.

Inception

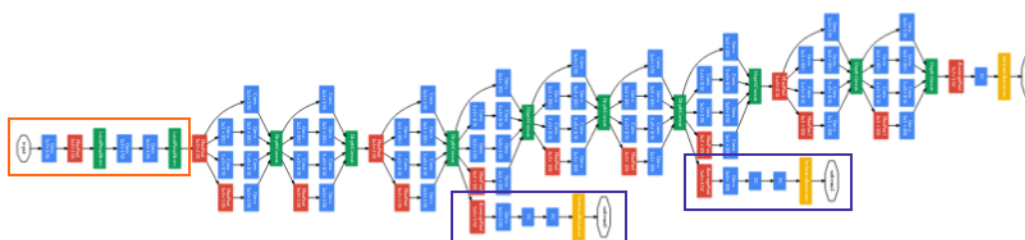


Figure 2.6: GoogNet or Inception Architecture

The winner of ILSVRC 2014 was GoogleNet or Inception networks [74]. The idea behind Inception was to use filters of different sizes (for example 3×3 , 5×5 , 7×7) parallelly, and thereby process the input at different resolutions and finally concatenate them before passing to the next layer. Their network was deeper than AlexNet and had auxiliary classifiers attached to the intermediate layers in order to boost the main classifier.

ResNet

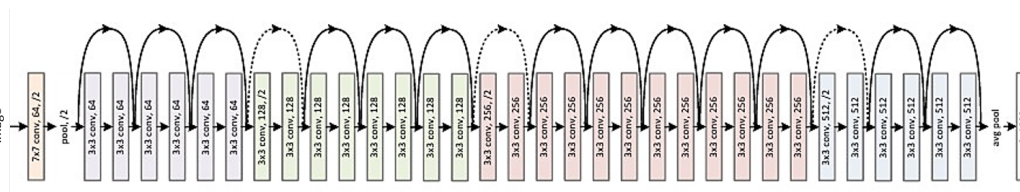


Figure 2.7: ResNet Architecture

Residual Networks or ResNet by He *et al.* [5] won the ILSVRC 2015 challenge. The idea behind ResNet was to increase the number of layers significantly and add residual connections to avoid vanishing gradients. They proposed several versions of the network using 18, 50, 101 and 152 layers. In

ResNet, the input to a certain layer is added to the output of that layer before passing it to the next layer. The authors also add computationally efficient bottleneck layers which let them design a very deep network with less parameters.

DenseNet

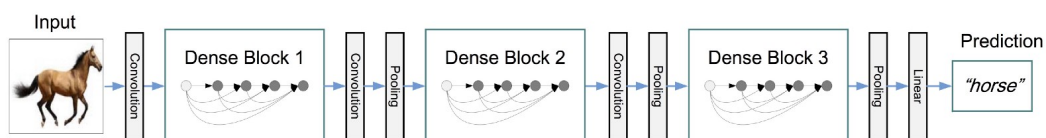


Figure 2.8: DenseNet Architecture

Huang *et al.* [75] proposed DenseNet which improved ResNet using denser connections but with fewer parameters. While in ResNet information from the last layer is passed using element-wise addition, DenseNet combines information from *all* the previous layers using concatenation. The reduction in the number of parameters is achieved by reducing the number of output channels. While in ResNet, the number of output channels is increased twice or four times the previous layer, in DenseNet the increase is mostly due to concatenation of channels from the previous layers.

2.3.3 Vision and Language

The work presented in Chapter 4 draws motivation from existing research in multi-modal problems that combine image and text such as natural image captioning (NIC). Given an input image, the task is to generate captions or textual descriptions of the scene. The captions describe the physical properties of the objects in the scene and the semantic relationship between them such as “a person riding a horse” or “a red car in front of a white building” etc.

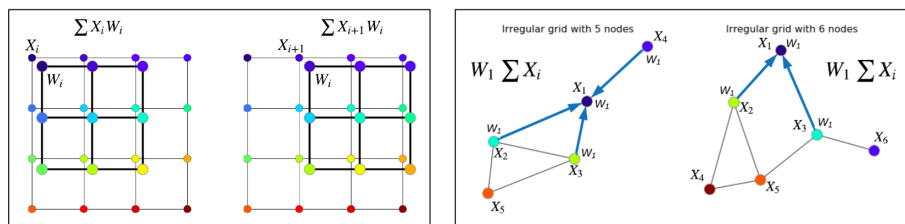
Early captioning methods typically follow a dictionary lookup approach [76, 77, 78, 79, 80]. Hodosh *et al.* [76] compiled the Flickr 8K dataset, comprising of 8000 images with 5 captions per image, which was extensively used in later works. They reported benchmark results using multiple methods based on minimally supervised features. They also conducted experiments to compare automatic evaluation metrics and human judgements. Socher *et al.* [77] propose a system which can retrieve an image based on a textual query and vice versa. They propose an RNN based encoding for the text which extracts meaningful components from the text by focusing on actions and agents in a sentence. Using markov random fields, Farhadi *et al.* [78] define a “meaning space” of triplets <object, action, scene> and project images and text to this space for comparing similarity. They also present a new dataset for the task. Ordonez *et al.* [79] compile the SBU Captioned Photo Dataset with 1 million images with associated text descriptions. They exploit global descriptors such as GIST and also a structure aware resized smaller version of the image for a bag of visual words based retrieval system. Jia *et al.* [80] propose a cross modal learning strategy Markov random field and topic modelling techniques.

More recent deep learning based methods prefer a parametric approach and are generative in the sense that they learn a mapping from visual to textual modality. Typically in these frameworks, a CNN is followed by a RNN or LSTM [81, 82, 83, 84, 85, 86, 87, 88, 89], although fully convolutional systems have been proposed by Aneja *et al.* [9] recently. Xu *et al.* in their seminal work [81] *Show, Attend and Tell* use visual attention to train a recurrent network for generating image descriptions from visual features extracted using a CNN. Johnson *et al.* [82] propose a fully con-

volutional localization network that jointly localizes and describes salient regions in the image. Fang *et al.* [83] propose a three stage pipeline in which they use multiple instance learning to train visual object detectors, generate sentences from the detected word using language models and finally rank those sentences. Karpathy *et al.* [84] propose a structured objective for multimodal embedding space in which visual features from a CNN are aligned with language features from a bidirectional RNN. Mao *et al.* present a similar approach using a CNN-RNN pipeline [87] and follow up with subsequent works on novel concept discovery and description ambiguity [89, 85] and achieve state-of-the-art results on the MS-COCO captioning benchmark. Anne *et al.* propose a captioning method which does not require a paired image-caption dataset. They achieve this by leveraging large object recognition datasets and external text corpora and by transferring knowledge between semantically similar concepts. Leveraging the developments in Image Captioning, Donahue *et al.* [88] present a method for video captioning. With the maturation of traditional image captioning, in recent years, there has been a growing interest towards newer forms of captioning tasks such as stylized captioning [90], visual storytelling [91, 92], aesthetic image captioning [10] *etc.*

2.3.4 Learning from Noisy Data

Data dependency of very deep neural nets and the high cost of human supervision has led to a natural interest towards exploring the easily available web-based big data. Our weakly supervised method presented in Chapter 4 draws motivation from this area as well. Berg *et al.* [93] build a visual attribute vocabulary by jointly mining noisy text and image data from e-commerce images. NEIL [94] gathers information continuously from the internet and performs simultaneous labelling and knowledge extraction. In [95], Divvala *et al.* crawl through a vast number of online books and train a model to automatically discover diverse appearance attributes for a given concept. Sun *et al.* [96] discover discriminative visual concepts by parallelly training text and images from the web and apply it for retrieval. Vittayakorn *et al.* [97] use neural activations for automatic attribute discovery from noisy web data. Yashima *et al.* [98] use online e-commerce text to generate product descriptions.



Pixels can be seen as nodes in a graph having a regular grid like structure.

Graph convolutions can be defined similarly but additionally, one has to define an **adjacency matrix**

Figure 2.9: Standard convolutions vs graph convolutions

2.3.5 Graph Neural Networks

Our work presented in Chapter 5 is based on Graph Neural Networks (GNN). GNNs have recently become popular in computer vision due to their ability to process irregularly structured data and non-local information. We provide a visual analogy between standard convolutions and their graph counterpart in Figure 2.9. A formal definition is provided in Chapter 5. Pixels of an image can be considered as nodes of a graph where the graph *i.e.* the whole image has a fixed adjacency relation to its neighbouring nodes. In the context of convolution, the adjacency is defined by the size of the filter's receptive field. For GNNs this information is provided as an additional input and therefore each graph can have arbitrary adjacency relations. During a forward pass, a node is updated with information from its neighbours using a technique called message passing.

Existing literature can be broadly classified into spectral [99, 100, 101, 102] and non-spectral approaches [103, 104, 105, 106, 107, 108]. While the spectral methods operate in the Fourier domain, the non-spectral methods are suited for the spatial domain and have endless applications such as molecular property prediction [103, 106], 3D shape estimation from point clouds [104], *etc.* In a GNN, rich representations are learnt from an input graph by sharing information among neighbouring nodes. Several solutions have been proposed to handle the arbitrary graphs with a different degree for each node. For example, [107] learns weights for each degree and [108] samples a fixed-sized neighbourhood and aggregates them. [105] use self-attention to select nodes based on importance.

2.4 Summary

We conclude this chapter with a brief summary of the areas presented in the last few sections and highlighting how that connects with our work in the next three chapters.

In Section 2.1, we started this survey by looking back at the brief history of the evolution of photography. We discussed the role criticism plays in improving this craft and also the complexities involved, owing to the ambiguous and subjective nature of aesthetics properties. "*Automating the process of criticism using state-of-the-art methods in deep learning*" — is the key motivation behind this research. In Section 2.2, we looked at the classical and recent approaches for automated image aesthetic assessment. We also presented the details of the AVA dataset, the biggest publicly available benchmark for Image Aesthetics. In Section 2.3, we discuss the other related areas — recent advances in neural network architectures, multimodal problems in vision and language, learning from noisy/weakly supervised data and graph neural network architectures.

In the next three chapters, we choose to study three different applications in Image Aesthetics and choosing only three from this vast collection was tough. Our choice was based both on the popular and on the rarely explored but promising areas. For example the study on aesthetic attributes in Chapter 3 was motivated from [2, 1, 3]. On the other hand, the study on aesthetic score regression in Chapter 5 is the most popular topic in Image Aesthetics [56, 7]. However, the work on captioning in Chapter 4, is a fairly new area and we are aware of only one previous work [10]. But, we were intrigued by the large body of work in vision and language and found it was worth exploring its potential applications in Image Aesthetics.

Needless to say, many other different applications are possible. In fact (probably a cliché), the list of our unfinished or unsuccessful projects is much longer than this dissertation itself. We point to some of those directions in Chapter 6 which are left for future explorations. But, in the following chapters, we look at some of our successful projects.

Chapter 3

Geometry Aware Aesthetic Attribute Prediction

Photographs are characterized by different attributes like the Rule of Thirds, depth of field, vanishing-lines etc. The presence or absence of one or more of these attributes contributes to the overall artistic value of an image. In this chapter, we analyze the ability of deep learning based methods to learn such aesthetic attributes. We observe that although a standard CNN learns the texture and appearance based features reasonably well, its understanding of global and geometric features is limited by two factors. First, the data-augmentation strategies (cropping, warping, etc.) distort the composition of a photograph and affect the performance. Secondly, the CNN features, in principle, are translation-invariant and appearance-dependent. But some geometric attributes important for aesthetics, e.g. *the Rule of Thirds* (RoT), are position-dependent and appearance-invariant. Therefore, we propose a novel input representation which is geometry-sensitive, position-cognizant and appearance-invariant. We further introduce a two-column CNN architecture that performs better than the state-of-the-art in aesthetic attribute prediction. From our results, we observe that the proposed network learns both the geometric and appearance-based attributes better than the state-of-the-art .



Figure 3.1: **Output from our network**: A screenshot from our web-based application. Predicted attributes are shown with their probability values (one-vs-all). This is a shot from Majid Majidi's film 'The Colours of Paradise'. We see that rule of thirds (for child's position), shallow depth of field, complementary colours (green background and reddish foreground), image grain (because of the poor video quality) are all well identified.

3.1 Motivation

Understanding the attributes of a photographic composition are crucial both for capturing and appreciating a picture. Analyzing objectively, these attributes can be broadly categorized into local or appearance-based (focus, image-grain, etc.) and global or geometry-based (aspect ratio, RoT, framing, etc.). We discuss some of the popular attributes in Appendix B. In this chapter, we propose a system which models the content-oriented (objective) aesthetic attributes of a photograph. Motivated by the recent developments in CNNs, our system takes a photograph as an input and predicts the aesthetic attributes, as illustrated in Figure 3.1. There are several applications of automatic photographic style classification. For example, post-processing images and videos, tagging, organizing and mining large collections of photos for artistic, cultural and historical purposes, scene understanding, building assistive-technologies, content creation, cinematography, etc.

The traditional approach of using CNNs for natural image classification

is to forward a *transformed* version of the input through a series of convolutional, pooling and fully connected layers and obtain a classification score. The transformation is applied to create a uniform sized input for the network (crop, warp, etc.) or to increase variance of the input distribution (flip, change contrast, etc.) for better generalization on the test data [12]. Clearly, such traditional transformations fail to preserve the aesthetic attributes of photographs. For example, a random fixed-sized crop cannot capture the arrangement of subjects within the picture. On the other hand, although warping the input photograph to a fixed size preserves the global context of the subjects better than crop, it has three primary disadvantages. (i) Firstly, photographs possess different aspect ratios which is important for their geometric and aesthetic attributes. Warping every photograph to a fixed size, irrespective of their aspect ratios, distorts these properties. (ii) Secondly, warping interpolates the RGB planes which results in the loss of appearance-based attributes like depth of field or image-grain. (iii) Thirdly, for a small dataset, warping reduces the variation in the input data and causes overfitting. In [1], the authors mention that even a small CNN with three convolution layers overfit on AVA Style dataset [29] when warping is used as the augmentation strategy.

This calls for a representation which preserves both the appearance-based and geometry-based properties of a photograph and which generalizes well over test data. Multiple solutions to these problems have been proposed. In [1], Lu *et al.* propose a double column CNN architecture, where the first column accepts a cropped patch and the second column accepts a warped version of the entire input. In subsequent work [2], multiple patches are cropped from an input and forwarded through the network. The features from multiple patches are aggregated before the final fully-connected layer for classification. The authors argue that sending multiple patches from the same image encodes more global context than a single random crop. More recently, Ma *et al.* [27] follow a similar multiple-patch extraction approach, but the patches are selectively extracted based on saliency, pattern-diversity and overlap between the subjects. Essentially, these techniques attempt to incorporate global context into the features during a forward pass either by warping the whole input and sending it through an additional column or by providing multiple

patches from the input at the same time.

Although these traditional double column or multi-patch strategies improve the overall performance, we argue that these networks cannot properly learn the geometry of a photograph. It is because CNNs, in principle, are designed to be translation invariant [109, 110]. While they can learn how the subjects look like, they cannot capture whether the subjects are rightly positioned. Since the convolutional filters corresponding to a feature map share weights, they become translation-invariant and appearance-dependent. In other words, they are activated for an object irrespective of its location in the image. As a result, they fail to understand photographic attributes like RoT. One option to tackle this could be training a fully-connected network on the full images, but they have too many parameters and are hard to train.

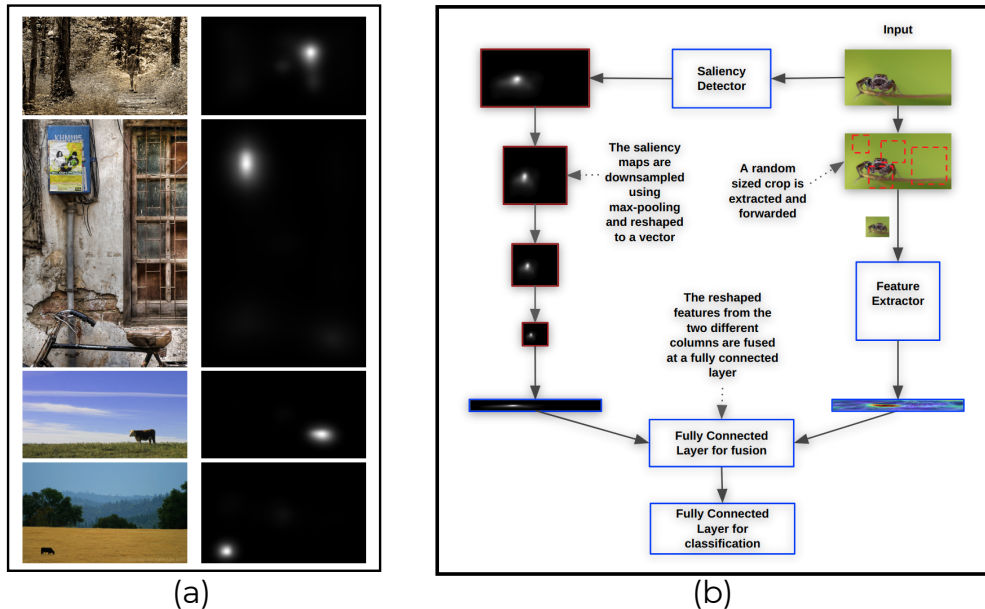


Figure 3.2: Our Contributions : **(a) Input (col 1), saliency maps (col 2)** : Saliency maps are generated using the method proposed in [8]. The position of the main subjects can be obtained from the saliency maps. **(b) Our double-column CNN architecture:** One column accepts the regular RGB features and the other column accepts saliency maps. The features from RGB channel are computed using a pre-trained Densenet161 [4], fine-tuned on our datasets. They are fused using a fully-connected layer and finally passed to another final fully-connected layer for classification.

In this context, we present two contributions in this chapter. The first one is introducing a saliency-based representations (see Figure 3.2(a)) which we call **Sal-RGB** features. The position or relative geometry of the different subjects in the image are obtained from the saliency maps and then fused with the appearance features coming from a traditional CNN and finally passed to a classifier to identify the overall style of composition of the photograph. By definition, saliency maps are appearance-invariant. On the other hand, by avoiding convolution and fusing them directly with the CNN features we achieve location-cognizance. In Section 3.4, we show that our approach performs better than the state-of-the-art in photographic style classification especially for those styles which are geometry-sensitive.

The second contribution is a comparative analysis of the traditional approaches for aesthetic categorization of images. Motivated both from the state-of-the-art and recent breakthroughs in deep learning, we implement multiple baselines, by trying different architectures and data augmentation strategies and try to understand and identify the factors that are crucial for encoding the local and global aspects of photographic composition.

Summary of contributions and outline

1. We propose a double-column convolutional neural network which fuses appearance and positional information and thereby addresses the limitations of a traditional CNN in handling geometric aesthetic attributes.
2. We thoroughly study the effects of different network architectures and augmentation strategies for the task of aesthetic attribute prediction.

The rest of the chapter is organized as follows. In Section 3.2, we describe the double column CNN architecture we adopt. In Section 3.3, we provide a detailed description of the datasets used. In Section 3.4, we provide details of the experiments conducted and analyze the results.

3.2 Network

In this section, we describe our architecture, as illustrated in Figure 3.2(b). Our architecture consists of three main blocks — the saliency detector [8], the double-column feature-extractor [4] and the classifier.

3.2.1 Saliency Detector

We compute the saliency maps using the method proposed in [8]. Motivated from recent attention based models [81, 111] that processes some regions of the input more attentively than others, the authors propose a CNN-LSTM framework for saliency detection. LSTMs are applied to sequential inputs where output from previous states are combined with inputs to the next state using dot products. In this work, the authors modify the standard LSTM such that they accept a sequence of spatial data (patches extracted from different locations in the image) and combine them using convolutions instead of dot products. Additionally, they introduce a center-prior component, that handles the tendency of humans to fix attention at the center region of an image. Some outputs from the system can be found in Figure 3.2(a), second column.

3.2.2 Feature Extractor

The feature extractor consists of two parallel and independent columns one for the saliency map and the other for raw RGB input.

Saliency Column : The saliency column consists of two max-pooling layers that downsample the input from 224×224 to 56×56 as shown in 3.2(b). Instead of max-pooling, we tried strided convolutions as they are known to capture low level details better than pooling [112, 113]. But pooling gave better results in our case which perhaps indicates that the salient position was more important than the level of detail captured.

RGB Column : We choose the DenseNet[16] [4] network for its superior performance in the ImageNet challenge. Very deep networks suffer from the *vanishing-gradient* problem *i.e.* gradual loss of information

as the input passes through several intermediate layers. Recent works like [75, 5, 114] address this problem by explicitly passing information between layers or by dropping random layers while training. The DenseNet is different from the traditional CNNs in the manner in which each layer receives input from the previous layers. The l^{th} layer in DenseNet receives as input, the concatenated output from all previous $l - 1$ layers. We provided additional details about the architecture in Chapter 2. We replace the last fully-connected layer from DenseNet with our classifier described in Section 3.2.3 and use the remaining as a feature extractor. Since we have less training images, we fine-tune [115] a model pre-trained on ImageNet on our dataset instead of training from scratch. This works since the lower level features like edges and corners are generic image features and can be used for aesthetic tasks too.

3.2.3 Classifier

Feature-maps from the two columns are concatenated and fused together using a fully-connected layer. A second and final fully-connected layer is used as a classifier. During training, we use the standard cross-entropy loss function and the gradient is back-propagated to the two columns.

3.2.4 Data Augmentation

As mentioned in Section 3.1, the training of a CNN is affected by how we provide the input data. It has been pointed out in [1] that while warping results in overfitting during training, cropping results in the loss of global context. We try five different data-augmentation techniques for the RGB column, to find the correlation of the input transformations and the different style attributes.

- **Centre Crop (T_{cc})** : Here, a patch of size $(s * s * 3)$ is extracted from the center of the image. This popular strategy is common in object-detection where the object of interest is centrally-localized.
- **Random crop of fixed size (T_{rc})** : Patches of size $(s*s*3)$ are extracted from the image at random locations. Although it suffers from data loss, it is effective in scenarios where the object of interest is not cen-

trally localized.

- **Warping (T_w)** : The input is anisotropically resized to a fixed dimension of $(s * s * 3)$. This transformation can squish objects along x or y. However it preserves global information better than a random crop.
- **Isotropic centre crop (T_{icc})** : In this case, the input is isotropically resized by warping its lower dimension to s . Then, a centre crop of $(s * s * 3)$ is applied. This preserves the geometric properties of the image at the cost of some global information.
- **Random crops of random sizes (T_{rnc})** : A crop of random size and random aspect ratio is made from the input. Then, it is resized to $(s * s * 3)$. We demonstrate in Section 3.4, that this strategy addresses the limitations of the previous strategies and performs best. We show a sample crop using this strategy in Figure 3.2(b).

In [1], the authors use T_{rc} for a single-column network and both T_{rc} and T_w for a double column network. In [2], the authors use T_{rc} .

It is to be noted here that we did not try appearance-based augmentation techniques like changing contrasts or colours since that would alter the style of the photograph.

3.3 Datasets

We use two standard datasets for evaluation — AVA Style and Flickr Style.

We already discussed AVA [29] in detail in Chapter 2. For our experiments we use AVA Style, which is a subset of AVA containing about 14,000 images with 14 aesthetic attributes. While training images in the subset are annotated with a single label, the test images have multiple labels associated with them making them unsuitable for popular evaluation frameworks used for single-label multi-class classifiers.

Flickr Style [3] is a collection of 80,000 images of 20 visual styles. The styles span across multiple concepts such as optical techniques (Macro, Bokeh, etc.), atmosphere (Hazy, Sunny, etc.), mood (Serene, Melancholy, etc.), composition styles (Minimal, Geometric, etc.), colour (Pastel, Bright, etc.) and genre (Noir,

Romantic, etc.). Flickr Style is a more complex dataset than AVA not only because it has more classes, but because some of the classes like Horror, Romantic and Serene are subjective concepts and difficult to encode objectively.

3.4 Experiments

We investigate two different aspects of the problem. First, in Section 3.4.1 we report the overall performance of our features using mean average precision (MAP). Second, in Section 3.4.2 we observe the per-class precision (PCP) scores to understand how our features affect individual photographic attributes. In Section 3.4.3, we analyze the effects of different data augmentation strategies described in Section 3.2.4. Finally, in Section 3.4.4, we discuss the limitations of our approach by analysing the misclassifications.

For comparison, we use MAP reported in [3, 1, 2]. PCP is compared only with [3] since the implementations were unavailable for [1, 2].

Additionally, we implement the following two benchmarks to evaluate our approach.

- **DenseNet161, ResNet152**: These are off-the-shelf implementations [4, 5] finetuned on our dataset and takes only RGB representation as input. These were chosen since they achieve the least error rates for ImageNet classification.
- **RAPID++**: Following [1], we implemented a two-column network. Each column takes as input random crops and the whole image as local and global representations, respectively. But, we used DenseNet161 architecture for the two columns whereas in the original work the authors use a shallower architecture with only three layers. We choose this as a benchmark in order to observe how their algorithm performs with a deeper architecture.

We train style classifiers on the AVA Style and Flickr Style datasets. The train-test partitions are followed from the original papers [29, 3]. For AVA, We use 11270 images for training and validation and 2573 images for testing. For Flickr Style we use 64000 images for training and 16000 images for

testing. The experiments are carried out on 2 NVidia Titan-XP GPUs. Each model was trained for 30 epochs with a learning rate of 0.0001. Training a model takes about 180 and 480 minutes for AVA and Flickr Style, respectively, with a batch-size of 16.

For testing, we follow the approach adopted by [1, 2]. 50 patches are extracted from the test-image and each patch is passed through the network. The results are averaged to achieve the final scores. Please note that this strategy does not affect the results for augmentation strategies like T_w , T_{icc} and T_{cc} . However, the results significantly improve for T_{rc} and T_{rnc} . The augmentation strategy is kept identical during training and testing.

3.4.1 Style Classification

Table 3.1: **Style Classification : Comparison with the state-of-the-art** : The results are reported in terms of Mean Average Precision(average of per class precision). We observe that for both the datasets, our method performs better than the state-of-the-art . Flickr Style was not used in [1, 2].

Network	Augmentation	AVA	Flickr Style
Fusion [3]	T_{cc}	58.10	36.80
RAPID [1]	T_{rc}, T_w	56.81	-
Multi-Patch [2]	T_{rc}	64.07	-
DenseNet161 [4]	T_{rnc}	71.68	43.83
ResNet152 [5]	T_{rnc}	70.57	43.65
RAPID++	T_{rc}, T_w	70.48	41.93
Sal-RGB	T_{rnc}	71.82	43.45

The scores are reported in terms of Mean Average Precision (MAP). MAP refers to the average of per-class precision. The results are reported in Table 3.1. We observe that our method outperforms the state-of-the-art [3, 1, 2] significantly. But, our own baselines perform more or less equally well. We deduce that for the improvement of MAP, the maximum impact is made by a more sophisticated CNN, followed by the location specific saliency. Both ResNet [5] and DenseNet [4] are residual networks and learn complex representations due to their very deep architectures. Such representations are crucial for learning photographic attributes, which have

many overlapping properties (less inter-class variance).

From these results, one might argue that the improvement can be attributed largely to a better CNN, and so what does Sal-RGB bring to the representation? We address this issue in Section 3.4.2.

3.4.2 Per-class Precision Scores

In [3], the authors report per-class precision (PCP) scores on AVA Style and Flickr Style. We compare our algorithm with those results in Table 3.2 and 3.3. We observe that our method outperforms [3] in almost all categories on both datasets. For the AVA Style dataset, a significant improvement is observed in the appearance-based categories like complementary colours, duotones, image grain, etc. Yet again, our own baselines DenseNet, ResNet and RAPID++ perform equally well in most categories except for RoT. For this category, Sal-RGB outperforms all others by a significant margin. This is an important result, since unlike others, RoT is a purely geometric attribute and important for Image Aesthetics and photography. A significant improvement in this category is a confirmation of our claim that the proposed approach efficiently encodes the geometry of a photograph. We highlight these observations in Figure 3.3.

3.4.3 Effect of data-augmentation

In this section, we observe the effects of different data-augmentation strategies described in Section 3.2.4. We plot the training performance in Figure 3.4 and show the test results in Table 3.4. We find T_{rnc} to be the best strategy for the task. T_w , although preserves global information more than T_{rc} , results in overfitting. The results are consistent with the observations in [1]. We tried the same experiments on Flickr and the results were similar.

3.4.4 Limitations

We tried to understand the limitations of our approach by plotting the confusion matrix for the different attributes of AVA and Flickr.

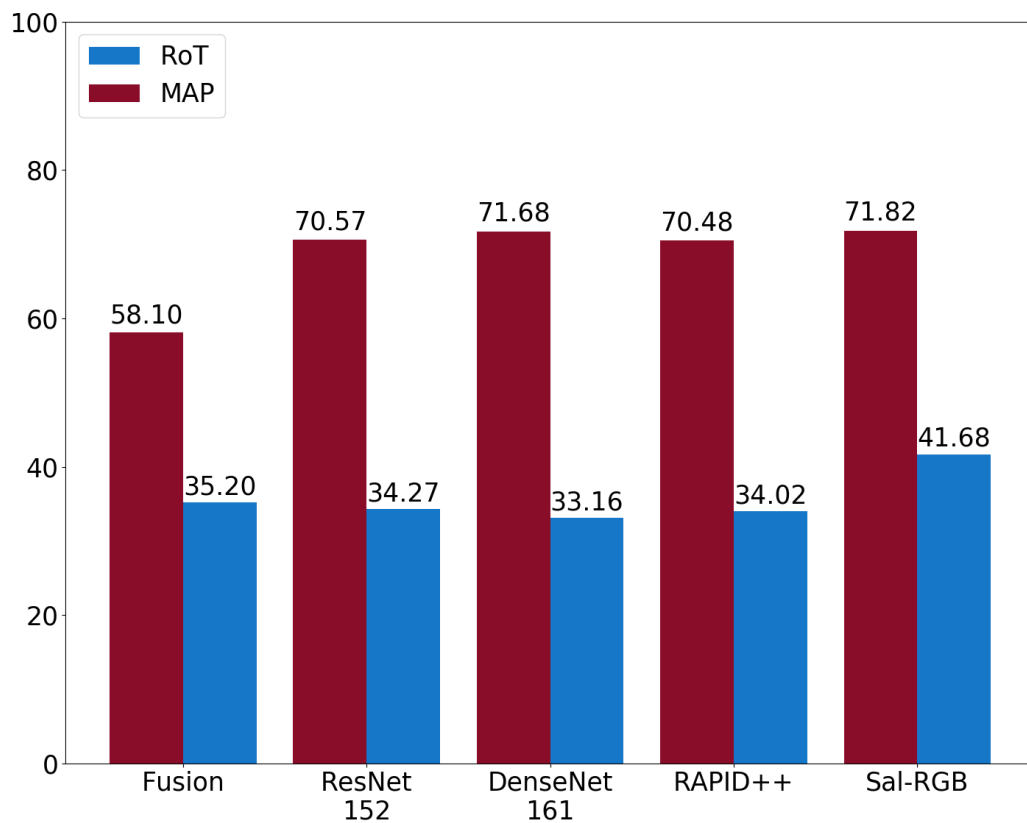


Figure 3.3: **Comparison of overall MAP and RoT precision for different networks:** We trained ResNet152 [5] and DenseNet161 [4] on AVA Style [5] and Fusion results are from [3]. RAPID++ is implemented following the data augmentation as done in [1] but with Densenet161 architecture. Although the MAP values are not too different, Sal-RGB outperforms others in finding RoT by a significant margin.

Table 3.2: **PCP for AVA Style Dataset** : Sal-RGB outperforms the state-of-the-art [3] by a significant margin in every category. Our own baselines DenseNet [4], ResNet [5], RAPID++ perform equally well for almost all categories except RoT, for which Sal-RGB performs much better.

Styles	Karayev (state-of-the-art) [3]	Densenet 161 [4]	ResNet 152 [5]	RAPID ++	Sal- RGB
Complementary_Colors	46.9	62.33	62.15	61.49	61.41
Duotones	67.6	86.58	84.82	84.77	87.58
HDR	66.9	74.95	70.08	71.51	72.86
Image_Grain	64.7	81.55	79.48	83.15	82.20
Light_On_White	90.8	84.69	83.41	85.64	82.99
Long_Exposure	45.3	64.16	65.38	63.94	61.94
Macro	47.8	64.89	65.52	64.90	66.58
Motion_Blur	47.8	63.93	62.12	61.21	61.98
Negative_Image	59.5	87.40	86.11	82.01	87.71
Rule_of_Thirds	35.2	33.16	34.27	34.02	41.68
Shallow_DOF	62.4	82.08	82.42	82.95	82.39
Silhouettes	79.1	93.73	92.49	91.14	93.05
Soft_Focus	31.2	49.89	44.91	44.57	46.41
Vanishing_Point	68.4	74.16	74.80	75.45	76.76

Table 3.3: **PCP for Flickr Style dataset** : Sal-RGB outperforms the state-of-the-art [3] by a significant margin. Our own baselines DenseNet [4], ResNet [5], RAPID++ perform equally well. The categories in Flickr are mostly appearance based. Hence, no significant improvement is achieved by using Sal-RGB over a regular CNN. Even the *geometric composition* category contain photographs of objects having regular geometric shapes . Hence, it is not location dependent in true sense.

Styles	Karayev (state-of-the-art) [3]	Densenet 161 [4]	ResNet 152 [5]	RAPID ++	Sal- RGB
Bokeh	28.80	30.24	31.34	29.39	29.78
Bright	25.10	22.97	23.12	22.69	23.33
Depth_of_Field	16.90	18.24	17.28	16.19	17.91
Detailed	33.70	37.96	38.27	38.50	38.09
Ethereal	40.80	50.31	50.88	48.15	50.03
Geometric_Composition	41.10	47.56	47.57	45.47	47.83
Hazy	48.70	61.59	60.01	57.68	60.92
HDR	49.30	65.44	65.24	61.03	64.92
Horror	40.00	64.24	64.17	58.40	64.16
Long_Exposure	51.50	65.36	64.76	61.40	63.62
Macro	61.70	67.44	70.26	69.60	68.18
Melancholy	16.80	19.82	20.33	18.50	19.71
Minimal	51.20	45.78	46.22	46.18	45.34
Noir	49.40	58.40	57.27	54.69	57.86
Pastel	25.80	34.15	34.05	30.71	34.17
Romantic	22.70	30.13	25.15	25.76	28.62
Serene	28.10	30.41	30.04	30.04	29.80
Sunny	50.00	59.99	60.56	58.57	58.58
Texture	26.50	28.98	30.52	29.72	29.65
Vintage	28.20	37.60	36.02	35.97	36.55

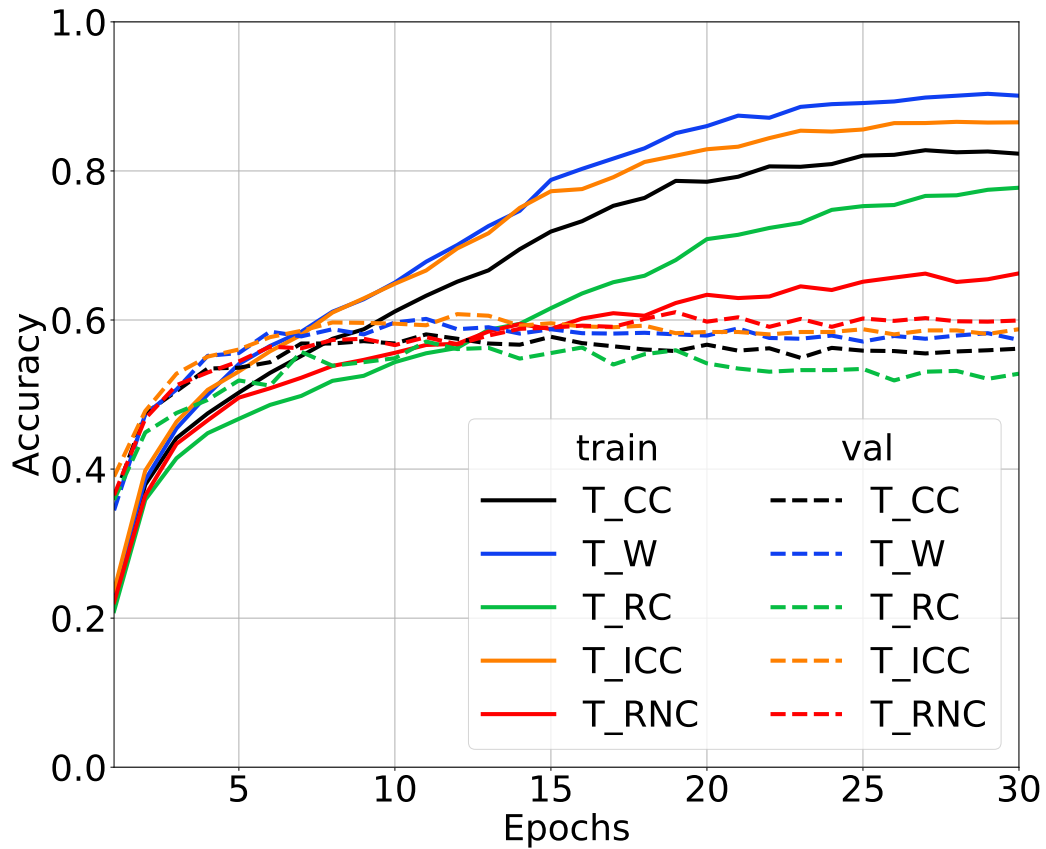


Figure 3.4: **Training and validation accuracy (normalized to [0,1]) for AVA Style:** For each strategy the proposed two-column network was trained for 30 epochs with a learning rate of 0.001 and a batch-size of 16. We observe, that in terms of overfitting (the gap between training and validation curves), the T_{rnc} and T_w performs best and worst, respectively. The decreasing order of overfitting is observed as follows $T_w > T_{icc} > T_{cc} > T_{rc} > T_{rnc}$. This observation is consistent with [1] where they observe that warping causes overfitting. In our case, both T_w and T_{icc} involve warping and hence are the most overfitted strategies.

Class	T_CC	T_W	T_RC	T_ICC	T_RNC	Fusion [3]
Complementary Colors	54.50	60.15	58.91	60.07	62.34	46.90
Duotones	82.32	84.48	85.64	84.33	84.91	67.60
HDR	55.05	61.95	70.54	61.67	72.12	66.90
Image_Grain	79.28	76.73	85.12	78.81	84.03	64.70
Light_On_White	77.73	86.04	85.76	84.10	86.59	90.80
Long_Exposure	53.96	62.65	58.90	59.23	61.87	45.30
Macro	62.06	63.21	68.54	63.48	67.29	47.80
Motion_Blur	54.84	57.56	64.19	58.53	63.69	47.80
Negative_Image	74.15	77.78	87.57	79.54	86.85	59.50
Rule_of_Third	39.02	41.06	36.30	42.02	39.39	35.20
Shallow_DOF	75.39	81.34	81.40	80.85	82.43	62.40
Silhouettes	88.34	90.87	92.21	92.06	93.48	79.10
Soft_Focus	36.60	34.94	46.08	35.77	48.18	31.20
Vanishing_Point	63.46	72.70	66.78	70.67	74.68	68.40
MAP	64.05	67.96	70.57	67.94	71.99	58.11

Table 3.4: **PCP and MAP for AVA with different augmentation** : We observe that a better validation accuracy ensures a better test performance. The decreasing order of mean average precision is as follows : $T_{rnc} > T_w > T_{rc} > T_{icc} > T_{cc}$

Confusions for AVA : The confusion matrix for AVA is plotted in Figure 3.5. Analysing the matrix, we observe the following

- The strongest classes are Light on White, Silhouettes, Vanishing Points. The weakest are Motion Blur and Soft Focus.
- Long Exposure and Motion Blur get confused with each other, which makes sense, since both attributes are captured using a slow shutter speed and mostly at night. (Figure. 3.7, row 1)
- Shallow DOF, Soft Focus and Macro are mutually confused classes, which is justified as all of them involve blur. (Figure. 3.7, row 2)
- The poorly performing classes have a high false-positive rate. We blame this on two factors. First, some classes such as Motion Blur and Soft-Focus have less samples as compared to others. Secondly, we observe that there is some ambiguity in the annotation of the training data of AVA. They are associated with a single label. But usually, most of the good photographs are captured with an interplay between multiple attributes. For example a macro image could very well conform to RoT or depth of field. Thus a single annotation incorporates undesired penalties to the loss during training the network and creates confusions during prediction.

Confusions for Flickr The confusion matrix for Flickr Style can be found in in Figure 3.6.

- The strongest classes are Long Exposure, HDR, Macro and Sunny. The weakest are Bright, DoF, Melancholy, Romantic and Serene.
- Melancholy, Romantic and Serene are subjective emotional properties. It can be argued that they need a lot of supervision to be effectively captured by a CNN.
- The Bright category of the dataset is diverse and hence is not properly learnt. It has many false positives. Hence the poor performance.
- There is a major mutual confusion between DoF and Bokeh, which makes sense, as both involve blur. (Figure. 3.7, row 3)
- Geometric Composition and Minimal composition are mutually con-

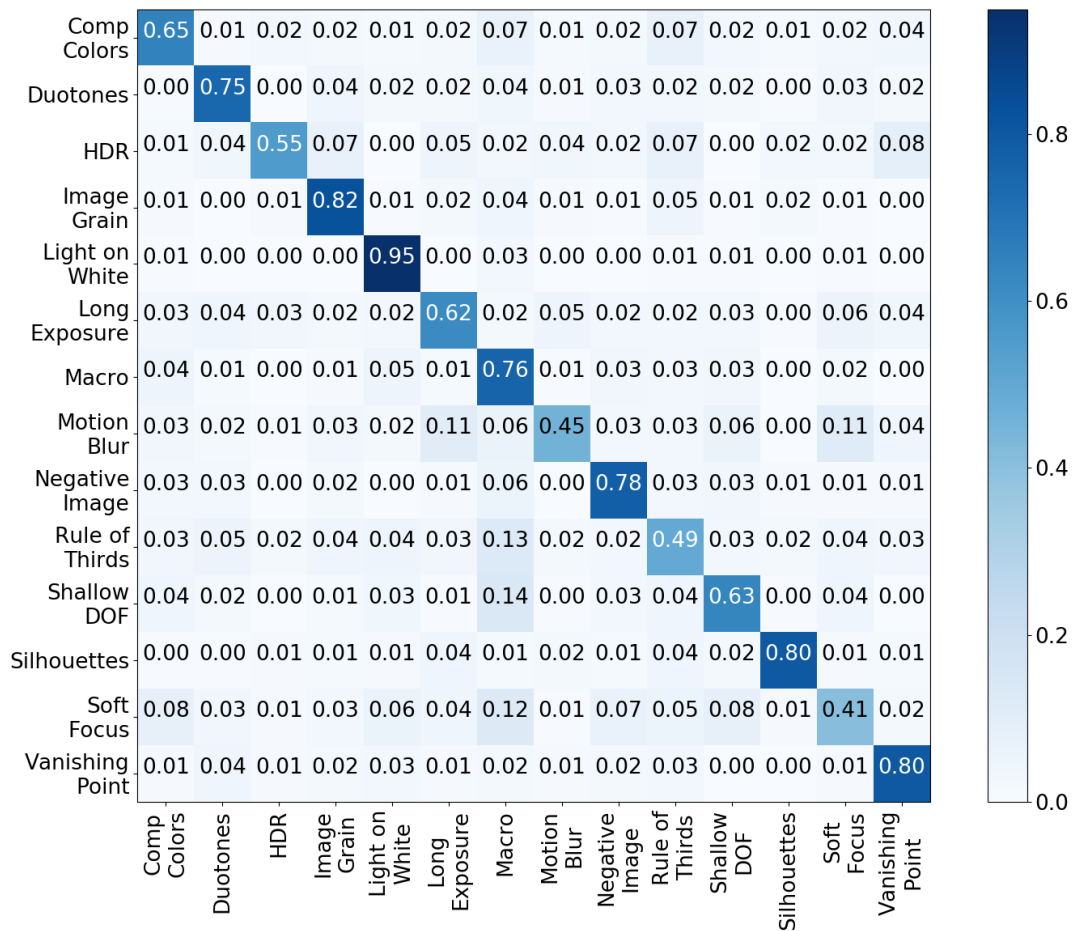


Figure 3.5: **Confusion matrix for AVA Style with our model:** For a test sample, the rows correspond to the real class and the columns correspond to the predicted class. The values are computed over 2573 test samples of AVA and then normalized. Examples of false positive images can be found in Figure 3.7

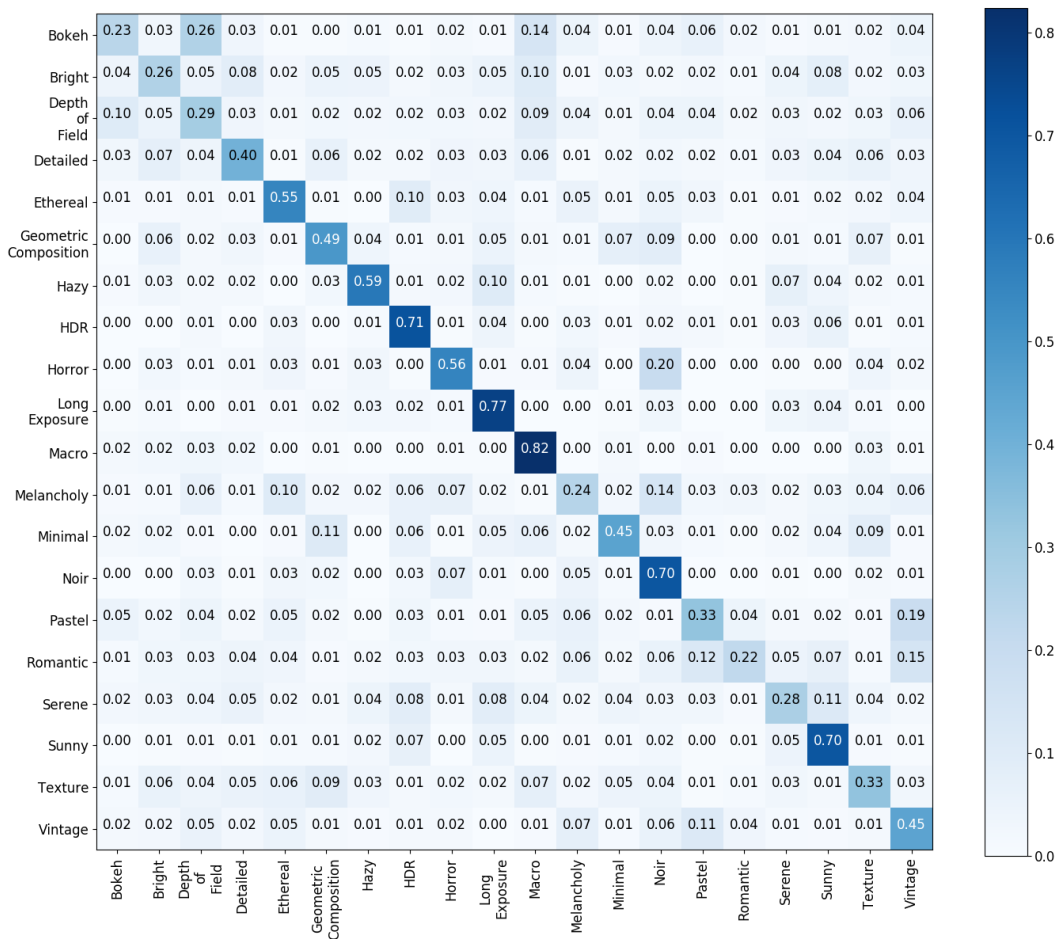


Figure 3.6: **Confusion Matrix for our model on Flickr Dataset**: Examples of false positive images can be found in Figure 3.7

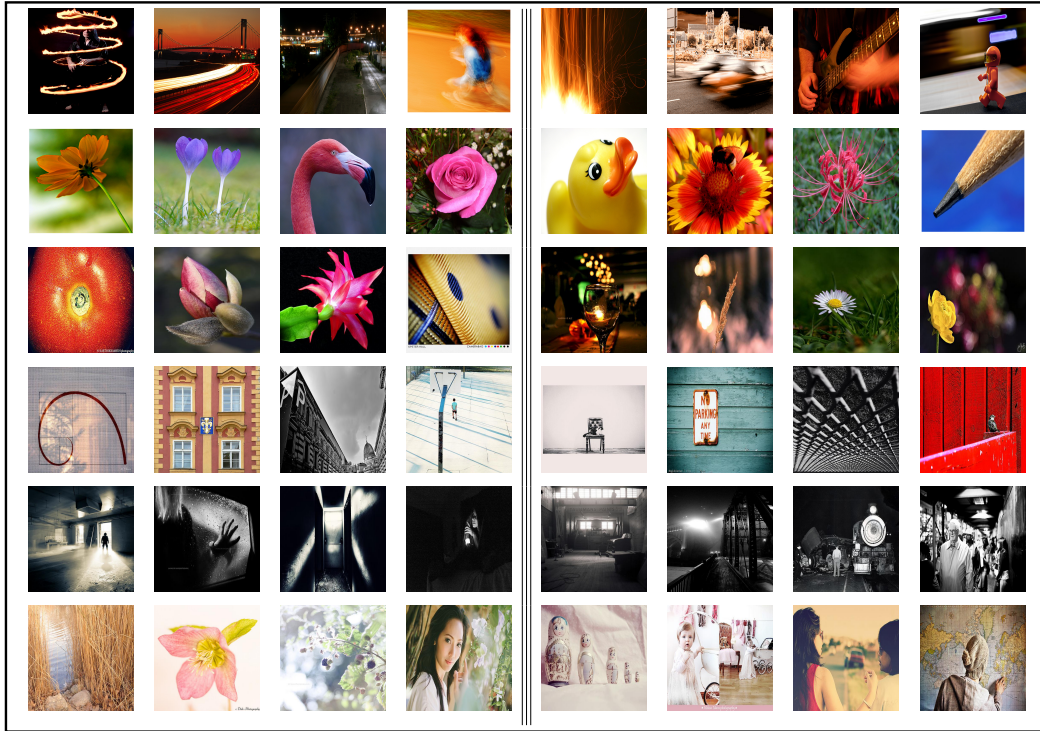


Figure 3.7: **False positives** : Each row corresponds to false positive samples from a pair of mutually confused classes. Column 1-4 and column 5-8 correspond to the first and second category in a pair, respectively. **Top-Bottom** - Long Exposure / Motion Blur, Shallow DOF / Macro, Shallow DOF / Bokeh, Geometric Composition / Minimal, Horror / Noir, Pastel / Vintage

fused. It makes sense, because both of them often involve lines and patterns. (Figure. 3.7, row 4)

- Horror and Noir are mutually confused since both contain photographs having dark appearances. (Figure. 3.7, row 5)
- Pastel and Vintage are mutually confused. Both the categories have a washed out/soothing appearance, so the confusions make sense. (Figure. 3.7, row 6)

3.5 Conclusion

In this work, we utilized the power of very deep neural networks and advanced the state-of-the-art in aesthetic attribute prediction. We intro-

duced a novel input representation which can be used with any state-of-the-art CNN architecture. Our method efficiently captures the geometry of a photograph, while preserving its local properties. We believe this representation can be applied to other domains where the relative positions of the objects are important (eg. scene classification). We conducted multiple experiments to understand the advantages and limitations of our approach and compared it with the state-of-the-art. There are many potential applications of an automatic style and aesthetic quality estimator in the domain of digital photography such as interactive cameras, automated photo correction *etc.* Our system can be directly extended to video-processing for predicting shot-styles. For example, Figure 3.1 illustrates the aesthetic analysis of a shot taken from Majid Majidi's movie *Colours of Paradise*.

As future work, there are many possible directions. Generalizing the model to more style attributes could be one. Extending the system to the domain of video and 360 images would also be possible. A thorough mathematical analysis of seemingly intangible and subjective concepts in art and subsequently fixing ambiguities in the data-annotation could be another. We hope that this area will become more active in the future with its challenging and interesting set of problems.

Chapter 4

Aesthetic Image Captioning from Weakly Labelled Photographs

Aesthetic image captioning (AIC) refers to the multi-modal task of generating critical textual feedbacks for photographs. While in natural image captioning (NIC), deep models are trained in an end-to-end manner using large curated datasets such as MS-COCO, no such large-scale, clean dataset exists for AIC. Towards this goal in this chapter, we propose an automatic cleaning strategy to create a benchmarking AIC dataset, by exploiting the images and noisy comments easily available from photography websites. We propose a probabilistic caption-filtering method for cleaning the noisy web-data, and compile a large-scale, clean dataset ‘AVA-Captions’, ($\sim 230,000$ images with ~ 5 captions per image). Additionally, by exploiting the latent associations between aesthetic attributes, we propose a strategy for training a CNN based visual feature extractor, typically the first component of an AIC framework. The strategy is weakly supervised and can be effectively used to learn rich aesthetic representations, without requiring expensive ground-truth annotations. We finally showcase a thorough analysis of the proposed contributions using automatic metrics and subjective evaluations.

4.1 Motivation

Availability of large curated datasets such as MS-COCO [50] (100K images), Flickr30K [116] (30K images) or Conceptual Captions [117] (3M images) made it possible to train deep learning models for complex, multi-modal tasks such as natural image captioning (NIC) [81] where the goal is to factually describe the image content. Similarly, several other captioning variants such as visual question answering [118], visual storytelling [91], stylized captioning [90] have also been explored. Recently, the PCCD dataset (~ 4200 images) [10] opened up a new area of research of describing images aesthetically. Aesthetic image captioning (AIC) has potential applications in the creative industries such as developing smarter cameras or web-based applications, ranking, retrieval of images and videos etc. However in [10], only six well-known photographic/aesthetic attributes such as composition, color, lighting, etc. have been used to generate aesthetic captions with a small curated dataset. Hence, curating a large-scale dataset to facilitate a more comprehensive and generalized understanding of aesthetic attributes remains an open problem.

Large-scale datasets have always been pivotal for research advancements in various fields [49, 50, 116, 119]. However, manually curating such a dataset for AIC is not only time consuming, but also difficult due to its subjective nature. Moreover, a lack of unanimously agreed ‘standard’ aesthetic attributes makes this problem even more challenging as compared to its NIC counterpart, where deep models are trained with known attributes/labels [50].

In this chapter, we make two contributions. Firstly, we propose an automatic cleaning strategy to generate a large scale dataset by utilizing the noisy comments or aesthetic feedback provided by users for images on the web. Secondly, for a CNN-based visual feature extractor as is typical in NIC pipelines, we propose a weakly-supervised training strategy. By automatically discovering certain ‘meaningful and complex aesthetic concepts’, beyond the classical concepts such as composition, color, lighting, etc., our strategy can be adopted in scenarios where finding clean ground-truth annotations is difficult (as in the case of many commercial applications). We elaborate these contributions in the rest of this section.





Training Strategy				
(a) Noisy Data & Supervised CNN (NS)	i like the angle and the composition	i like the colors and the composition	i like the composition and the lighting	i like the composition and the bw
(b) Clean Data & Supervised CNN (CS)	i like the idea , but i think it would have been better if the door was in focus .	i like the colors and the water . the water is a little distracting .	i like the way the light hits the face and the background .	i like this shot . i like the way the lines lead the eye into the photo .
(c) Clean Data & Weakly Supervised CNN (CWS)	i like the composition , but i think it would have been better if you could have gotten a little more of the building	i like the composition and the colors . the water is a little too bright .	this is a great shot . i love the way the light is coming from the left .	i like the composition and the bw conversion .

Figure 4.1: Aesthetic image captions. We show candidates generated by three different frameworks discussed in this chapter: **(a)** For NS, we use an ImageNet trained CNN and LSTM trained on noisy comments **(b)** For CS, we use an ImageNet trained CNN and LSTM trained on compiled AVA-Captions dataset **(c)** For CWS, we use a weakly-supervised CNN and LSTM trained on AVA-Captions

To generate a clean aesthetic captioning dataset, we collected the raw user comments from the Aesthetic Visual Analysis (AVA) dataset [29]. AVA is a widely used dataset for aesthetic image analysis tasks such as aesthetic rating prediction [1, 27], photographic style classification [3, 31]. However, AVA was not created for AIC. In this chapter, we refer to the original AVA with raw user comments as AVA raw-caption. It contains $\sim 250,000$ photographs from dpchallenge.com and the corresponding user comments or feedback for each photograph (3 billion in total). Typically, in Dpchallenge, users ranging from casual hobbyists to expert photographers provide feedback to the images submitted and describe the factors that make a photograph aesthetically pleasing or dull. Even though these captions contain crucial aesthetic-based information from images, they cannot be directly used for the task of AIC. Unlike the well instructed and curated datasets [50], AVA raw-captions are unconstrained user-comments in the wild with typos, grammatically inconsistent statements, and also containing a large number of comments occurring frequently without useful information. Previous work in AIC [10] acknowledges the difficulty of dealing with the highly noisy captions available in AVA.

In this work, we propose to clean the raw captions from AVA by proposing a probabilistic n-gram based filtering strategy. Based on word-composition and frequency of occurrence of n-grams, we propose to assign an informativeness score to each comment, where comments with a little or vague information are discarded. Our resulting clean dataset, **AVA-Captions** contains $\sim 230,000$ images and $\sim 1.5M$ captions with an average of ~ 5 comments per image and can be used to train the LSTM network in the image captioning pipeline in the traditional way. Our subjective study verifies that the proposed automatic strategy is consistent with human judgment regarding the informativeness of a caption. Our quantitative experiments and subjective studies also suggest that models trained on AVA-Captions are more diverse and accurate than those trained on the original noisy AVA-Comments. It is important to note that our strategy to choose the large-scale AVA raw-caption is motivated from the widely used image analysis benchmarking dataset, MS-COCO, which is now used as a unified benchmark for diverse tasks such as object detection, segmentation, captioning, etc. We hope that our cleaned dataset will serve as a new

benchmarking dataset for various creative studies and aesthetics-based applications such as aesthetics based image enhancement, smarter photography cameras, etc.

Our second contribution in this work is a weakly supervised approach for training a CNN, as an alternative to the standard practice. The standard approach for most image captioning pipelines is to train a CNN on large annotated datasets e.g. ImageNet [49], where rich and discriminative visual features are extracted corresponding to the physical properties of objects such as cars, dogs etc. These features are provided as input to an LSTM for generating captions. Although trained for classification, these ImageNet-based features have been shown to translate well to other tasks such as segmentation [120], style-transfer [121], NIC. In fact, due to the unavailability of large-scale, task-specific CNN annotations, these ImageNet features have been used for other variants of NIC such as aesthetic captioning [10], stylized captioning [90], product descriptions [98], etc.

However, for many commercial/practical applications, availability of such datasets or models is unclear due to copyright restrictions [122, 123, 124]. On the other hand, collecting task-specific manual annotations for a CNN is expensive and time intensive. Thus the question remains open if we can achieve better or at least comparable performance by utilizing easily available weak annotations from the web (as found in AVA) and use them for training the visual feature extractor in AIC. To this end, motivated from weakly supervised learning methods [125, 126], we propose a strategy which exploits the large pool of unstructured raw-comments from AVA and discovers latent structures corresponding to meaningful *photographic concepts* using Latent Dirichlet Allocation (LDA) [127]. We experimentally observe that the weakly-supervised approach is effective and its performance is comparable to the standard ImageNet trained supervised features. In essence, our contributions in this chapter are as follows:

Contributions

1. We propose a caption filtering strategy and compile **AVA-Captions**, a large-scale and clean dataset for aesthetic image captioning (Sec 4.2).
2. We propose a **weakly-supervised approach** for training the CNN of a standard CNN-LSTM framework (Sec 4.3)
3. We showcase the analysis of the AIC pipeline based on the standard automated metrics (such as BLEU, CIDEr, SPICE etc. [128, 129, 130]), diversity of captions and **subjective evaluations** which are publicly available for further explorations (Section 4.5).

4.2 Caption Filtering Strategy

Image	Comments	Scores
	Photo Quality : Awesome	9.62
	I love the colors here	1.85
	I like the trees in the back-ground and the softness of the water .	28.41
	The post processing looks great with the water , but the top half of the photo doesn't work as well.	47.44

Figure 4.2: Informativeness of captions.

In AVA raw-caption, we observe two main types of undesirable captions. First, there are captions which suffer from generic noise frequently observed in most text corpora, especially those compiled from social media. They include typing errors, non-English comments, colloquial acronyms, exclamatory words (such as “woooow”), extra punctuation (such as “!!!!”), etc. Such noise can be handled using standard natural language processing techniques [131].

Second, we refer to the *safe* comments, which carry a little or no useful in-

formation about the photograph. For example, in Figure 4.2, comments such as “*Photo Quality : Awesome*” or “*I love the colors here*” provide a valid but less informative description of the photograph . It is important to filter these comments, otherwise the network ends up learning these less-informative, *safe* captions by ignoring the more informative and discriminative ones such as “*The post processing looks great with the water, but the top half of the photo doesn’t work as well.*” [10].

To this end, we propose a probabilistic strategy for caption filtering based on the informativeness of a caption. Informativeness is measured by the presence of certain n-grams. The approach draws motivation from two techniques frequently used in vision-language problems — word composition and term-frequency - inverse document frequency (TF-IDF).

Word Composition: Bigrams of the “descriptor-object” form often convey more information than the unigrams of the objects alone. For example, “post processing” or “top half” convey more information than “processing” or “half”. On the other hand, the descriptors alone may not always be sufficient to describe a complete concept and its meaning is often closely tied to the object [132]. For example, “sharp” could be used in two entirely different contexts such as “sharp contrast” and “sharp eyes”. This pattern is also observed in the 200 bigrams (or ugly and beautiful attributes) discovered from AVA by Marchesotti *et al.* [29] such as “nice colors”, “beautiful scene”, “too small”, “distracting background”, etc. Similar n-gram modelling is found in natural language processing as adjective-noun [77, 133, 134] or verb-object [135, 136] compositions.

TF-IDF: The other motivation is based on the intuition that the key information in a comment is stored in certain n-grams which occur less frequently in the comment corpus such as “softness”, “post processing”, “top half” etc. A sentence composed of frequently occurring n-grams such as “colors” or “awesome” is less likely to contain useful information. The intuition follows from the motivation of commonly used TF-IDF metric in document classification, which states that more frequent words of a vocabulary are less discriminative for document classification [137]. Such hypothesis also forms a basis in the CIDEr evaluation metric [129] widely used for tasks such as image captioning, machine translation, etc.

Proposed “Informativeness” Score: Based on these two criteria, we start by constructing two vocabularies as follows: for unigrams we choose only the nouns and for bigrams we select “descriptor-object” patterns *i.e.* where the first term is a noun, adjective or adverb and the second term is a noun or an adjective. Each n-gram ω is assigned a corpus probability P as:

$$P(\omega) = \frac{C_\omega}{\sum_{i=1}^D C_i} \quad (4.1)$$

where the denominator sums the frequency of each n-gram ω such that $\sum_{i=1}^D P(\omega_i) = 1$, where D is the vocabulary size, and C_ω is the corpus frequency of n-gram ω . Corpus frequency of an n-gram refers to the number of times it occurs in the comments from all the images combined. This formulation assigns high probabilities for more frequent words in the comment corpus.

Then, we represent a comment as the union of its unigrams and bigrams *i.e.*, $S = (S_u \cup S_b)$, where $S_u = (u_1 u_2 \dots u_N)$ and $S_b = (b_1 b_2 \dots b_M)$ are the sequences of unigrams and bigrams, respectively. A comment is assigned an informativeness score ρ as follows:

$$\rho_s = -\frac{1}{2} [\log \prod_i^N P(u_i) + \log \prod_j^M P(b_j)] \quad (4.2)$$

where $P(u)$ and $P(b)$ are the probabilities of a unigram or bigram given by Equation 4.1. Equation 4.2 is the average of the negative log probabilities of S_u and S_b .

Essentially, the score of a comment is modelled as the joint probability of n-grams in it, following the simplest Markov assumption *i.e.* all n-grams are independent [138]. If the n-grams in a sentence have higher corpus probabilities then the corresponding score ρ is low due to the negative logarithm, and vice-versa.

Note that the score is the negative logarithm of the product of probabilities and longer captions tend to receive higher scores. However, our approach does not *always* favour long comments, but does so only if they consist of “meaningful” n-grams conforming to the “descriptor-object” com-

position. In other words, randomly long sentences without useful information are discarded. On the other hand, long and informative comments are kept. This is also desirable as longer comments in AVA tend to be richer in information as expert users are specifically asked to provide detailed assessment which is referred to as *critique club effect* in [139].

We label a comment as informative or less-informative by thresholding (experimentally kept 20) the score ρ . Some sample scores are provided in Figure 4.2. The proposed strategy discards about 1.5M (55%) comments from the entire corpus. Subsequently, we remove the images which are left with no informative comments. Finally, we are left with 240,060 images and 1,318,359 comments, with an average of 5.58 comments per image. We call this cleaner subset as **AVA-Captions**. The proposed approach is evaluated by human subjects and the results are discussed in Figure 4.6 and Section 4.5.3.

4.3 Weakly Supervised CNN

Although the comments in AVA-Captions are cleaner than the raw comments, they cannot be directly used for training the CNN *i.e.* the visual feature extractor. As discussed in Sec 4.1, the standard approach followed in NIC and its variants is to use an ImageNet trained model for the task. In this section, we propose an alternative weakly supervised strategy for training the CNN from scratch by exploiting the *latent* aesthetic information within the AVA-Captions. Our approach is motivated from two different areas: visual attribute learning and text document clustering.

4.3.1 Visual and Aesthetic Attributes

Visual Attribute Learning is an active and well-studied problem in computer vision. Instead of high-level object/scene annotations, models are trained for low-level attributes such as “smiling face”, “open mouth”, “full sleeve” *etc.* and the features are used for tasks such as image-ranking [140], pose-estimation [141], fashion retrieval [142], zero-shot learning [143], *etc.* Similarly, our goal is to identify aesthetic attributes and train a CNN. A


Topics	Images
"Cute-Expression", "Face", "Ear"	
"Landscape", "Sky", "Cloud"	
"Action Shot", "Sport", "Great Action"	
"Black and white", "Tone", "Contrast"	
"Motion Blur", "Movement", "Shutter Speed"	

Figure 4.3: Some topics / labels discovered from AVA-Captions using LDA.

straightforward approach is to use the n-grams from comments (Sec 4.2) and use them as aesthetic attributes. However, there are two problems with this approach: Firstly, the set of n-grams is huge ($\sim 25K$) and thus training the CNN directly using them as labels is not scalable. Secondly, several n-grams such as "grayscale", "black and white", "bw" refer to the same concept and carry redundant information.

Therefore, we apply a clustering of semantically similar n-grams and thereby grouping the images which share similar n-grams in their corresponding comments. For example, portraits are more likely to contain attributes such as "cute expression", "face" etc. and landscape shots are more likely to share attributes such as "tilted horizon", "sky", "overexposed clouds" etc. Essentially, the intuition behind our approach is to discover clusters of photographic attributes or topics from the comment corpus and use them as labels for training the CNN. In text document analysis, it is a common practice to achieve such grouping of topics from a text corpus using a technique called Latent Dirichlet Allocation [127].

4.3.2 Latent Dirichlet Allocation (LDA)

LDA is an unsupervised generative probabilistic model, widely used for topic modelling in text corpora. It represents text documents as a probabilistic mixture of topics, and each topic as a probabilistic mixture of words. The words which co-occur frequently in the corpus are grouped together by LDA to form a topic. For example, by running LDA on a large corpus of news articles, it is possible to discover topics such as “sports”, “government policies”, “terrorism” etc [144].

Formally stated, given a set of documents $D_i = \{D_1, D_2 \dots D_N\}$, and a vocabulary of words $\omega_i = \{\omega_1, \omega_2 \dots \omega_M\}$, the task is to infer K latent topics $T_i = \{T_1, T_2, \dots T_K\}$, where each topic can be represented as a collection of words (term-topic matrix) and each document can be represented as a collection of topics (document-topic matrix). The term-topic matrix represents the probabilities of each word associated with a topic and the document-topic matrix refers to the distribution of a document over the K latent topics. The inference is achieved using a variational Bayes approximation [127] or Gibb’s sampling [145]. A more detailed explanation can be found in [127].

4.3.3 Relabelling AVA Images

We regard all the comments corresponding to a given image as a document. The vocabulary is constructed by combining the unigrams and bigrams extracted from the AVA-Captions as described in Section 4.2. In our case: $N = 230,698$ and $M = 25,000$, and K is experimentally fixed to 200. By running LDA with these parameters on AVA-Captions, we discover 200 latent topics, composed of n-grams which co-occur frequently. The method is based on the assumption that the visual aesthetic attributes in the image are correlated with the corresponding comments and images possessing similar aesthetic properties are described using similar words.

Even after the caption cleaning procedure, we observe that n-grams such as “nice composition” or “great shot” still occur more frequently than others. But, they occur mostly as independent clauses in bigger comments such as *“I like the way how the lines lead the eyes to the subject. Nice*

shot!". In order to avoid inferring topics consisting of these less discriminative words, we consider only those n-grams in the vocabulary which occur in less than 10% comments.

In Figure 4.3, we select 5 topics thus inferred and some of the corresponding images whose captions belong to these topics. It can be observed that the images and the words corresponding to each topic are fairly consistent and suitable to be used as labels for training the CNN.

4.3.4 Training the CNN

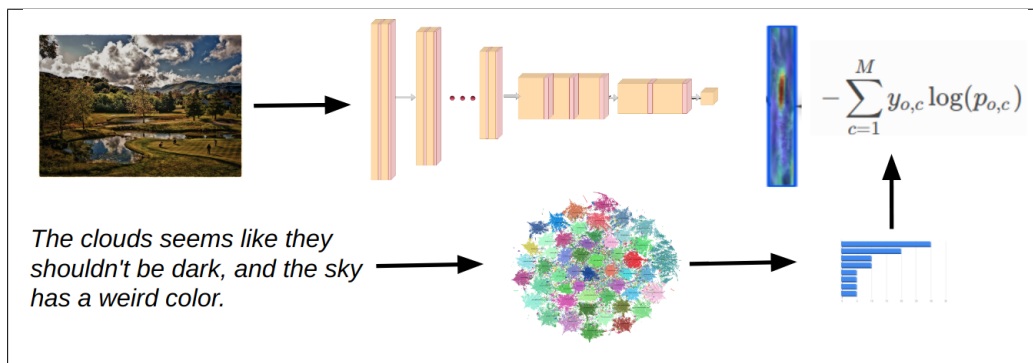
Given an image and its corresponding captions, we estimate the topic distribution D_T of the comments. The CNN is trained using D_T as the ground-truth label. We adopt the ResNet101 [5] architecture and replace the last fully connected layer with K outputs, and train the framework using cross-entropy loss [146] as shown in Figure 4.4a.

4.4 The Final Framework

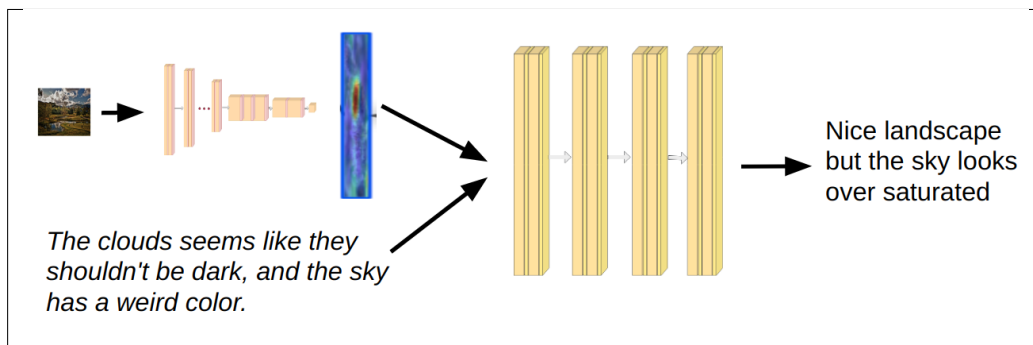
We adopt the NeuralTalk2 [147] framework as our basis. Note, that our approach is generic and can be used with any CNN-LSTM framework for image captioning. In [147], visual features are extracted using an ImageNet trained ResNet101 [5] which are passed as input to an LSTM for training the language model using the ground-truth captions. For our framework, we use two alternatives for visual features (a) ImageNet trained (b) weakly supervised (Sec 4.3). The LSTM architecture is kept unchanged except hyper-parameters such as vocabulary size, maximum allowed length of a caption *etc.* The language model is trained using the clean and informative comments from the AVA-Captions dataset (See Figure 4.4b).

4.5 Experiments

The experiments are designed to evaluate the two primary contributions: First, the caption cleaning strategy and second, the weakly-supervised training of the CNN. Specifically, we investigate: **(a)** the effect of caption filtering and the weakly supervised approach on the quality of captions



(a) Weakly-supervised training of the CNN: Images and comments are provided as input. The image is fed to the CNN and the comment is fed to the inferred topic model. The topic model predicts a distribution over the topics which is used as a label for computing the loss for the CNN.



(b) Training the LSTM: Visual features extracted using the CNN and the comment is fed as an input to the LSTM which predicts a candidate caption.

Figure 4.4: **Proposed pipeline**

generated in terms of accuracy (Sec 4.5.3) and diversity (Sec 4.5.3), **(b)** the generalizability of the captions learnt from AVA, when tested on other image-caption datasets (Sec 4.5.3), **(c)** subjective or human opinion about the performance of the proposed framework (Sec 4.5.3).

4.5.1 Datasets

AVA-Captions: The compiled AVA-Captions dataset is discussed in detail in Section 4.2. We use 230,698 images and 1,318,359 comments for training; and 9,362 images for validation.

AVA raw-caption: The original AVA dataset provided by Murray *et al.* [29] and the raw unfiltered comments are used to train the framework in order to observe the effects of caption filtering.

Photo Critique Captioning Dataset (PCCD): This dataset was introduced by [10] and is based on www.gurushots.com. Professional photographers provide comments for the uploaded photos on seven aspects: general impression, composition and perspective, color and lighting, subject of photo, depth of field, focus and use of camera, exposure and speed. In order to verify whether the proposed framework can generate aesthetic captions for images beyond the AVA dataset we trained it with AVA-Captions and tested it with PCCD. For a fair comparison, we use the same validation set provided in the original paper.

4.5.2 Baselines

We compare three implementations: **(a) Noisy - Supervised (NS):** NeuralTalk2 [147] framework trained on AVA-Original. It has an ImageNet trained CNN, followed by LSTM trained on raw, unfiltered AVA comments. NeuralTalk2 is also used as a baseline for AIC in [10]. **(b) Clean - Supervised (CS):** The LSTM of the NeuralTalk2 is trained on AVA-Captions *i.e.* filtered comments. The CNN is same as NS *i.e.* Imagenet trained. **(c) Clean and weakly-supervised (CWS):** NeuralTalk2 framework, where the CNN is trained with weak-supervision using LDA and the language model is trained on AVA-Captions.

Method	B1	B2	B3	B4	M	R	C	S	S-1
NS	0.379	0.219	0.122	0.061	0.079	0.233	0.038	0.044	0.135
CS	0.500	0.280	0.149	0.073	0.105	0.253	0.060	0.062	0.144
CWS	0.535	0.282	0.150	0.074	0.107	0.254	0.059	0.061	0.144

(a) Accuracy

Method	Train	Val	S-1	Precision	Recall
CNN-LSTM-WD	PCCD	PCCD	0.136	0.181	0.156
AO	PCCD	PCCD	0.127	0.201	0.121
AF	PCCD	PCCD	0.150	0.212	0.157
CS	AVA-C	PCCD	0.144	0.166	0.166
CWS	AVA-C	PCCD	0.142	0.162	0.161

(b) Generalizability

Table 4.1: (a) **Results on AVA-Captions:** Both CS and CWS, trained on AVA-Captions perform significantly better than NS, which is trained on noisy data. Also, the performance of CWS and CS is comparable, which proves the effectiveness of the weakly supervised approach (b) **Generalization results on PCCD:** Models trained on AVA-C perform well on PCCD validation set, when compared with models trained on PCCD directly. We argue that this impressive generalizability is achieved by training on a larger and diverse dataset.

4.5.3 Results and Analysis

Accuracy

Most of the existing standards for evaluating image captioning such as BLEU (B) [128], METEOR (M) [148], ROGUE (R) [149], CIDEr (C) [129] etc. are mainly more accurate extensions of the brute-force method [150] *i.e.* comparing the n-gram overlap between candidate and reference captions. Recently introduced metric SPICE (S) [130] instead compares scene graphs computed from the candidate and reference captions. It has been shown that SPICE captures semantic similarity better and is closer to human judgement more than the rest. Traditionally, SPICE is computed between the candidate and all the reference captions. A variant of SPICE (which we refer to as S-1) is used in [10] where the authors compute SPICE between the candidate and each of the reference captions and choose the best. In this

thesis, we report both S and S-1.

From Table 4.1(a), we observe that both CS and CWS outperform NS significantly over all metrics. Clearly, training the framework with cleaner captions results in more accurate outputs. On the other hand, the performance of CWS and CS is comparable. We argue that this indicates that the proposed weakly-supervised training strategy is capable of training the CNN as efficiently as a purely supervised approach and extract meaningful aesthetic features. Additionally as mentioned in Sec 4.1, the proposed CWS approach has an advantage that it does not require expensive human annotations to train. Thus, it is possible to scale to deeper architectures, and thus learn more complex representations simply by crawling the vast, freely-available and weakly-labelled data from the web.

Diversity

Image Captioning pipelines often suffer from monotonicity of captions *i.e.* similar captions are generated for the validation images. This is attributed to the fact that the commonly used cross-entropy loss function trains the LSTM by reducing the entropy of the output word distribution and thus giving a *peaky* posterior probability distribution [9]. As mentioned earlier in Section 4.1, this is more pronounced in AIC due to the vast presence of the *easy* comments in the web. Diversity of the captions is usually measured by overlap between the candidate and the reference captions. We evaluate diversity following two state-of-the-art approaches [10, 9]. In [10], the authors define that two captions are different if the ratio of common words between them is smaller than a threshold (3% used in the paper). In [9], from the set of all the candidate captions, the authors compute the number of unique n-grams (1, 2, 4) at each position starting from the beginning up to position 13.

We plot diversity using [10] in Figure 4.5d. We compute using the alternative approach of [9] in Figure 4.5(a-c) but up to 25 positions since on an average the AVA captions are longer than the COCO captions. From both, we notice that diversity of NS is significantly lesser than CS or CWS. We observe that NS ends up generating a large number of “safe” captions such as “I like the composition and colours” or “nice shot” *etc.* We argue,

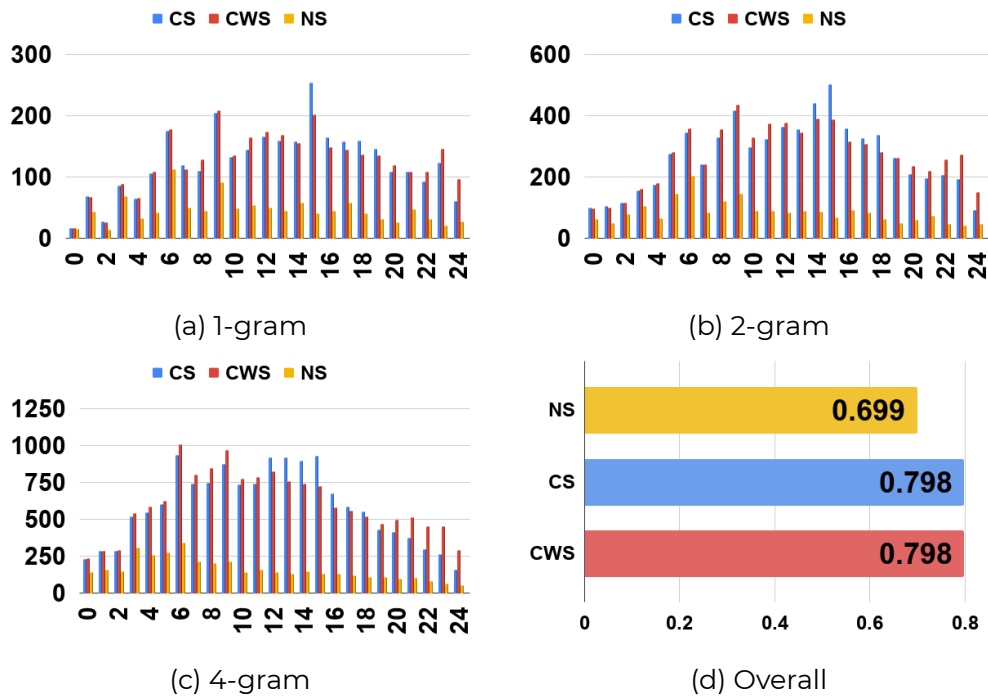


Figure 4.5: **Diversity:** Figures (a) - (c) report diversity of captions following [9]. The x -axes correspond to n -gram positions in a sentence. The y -axes correspond to the number of unique n -grams at each position, for the entire validation set. Figure (d) plots the overall diversity, as reported in [10]. We observe that the diversity of the captions increase significantly when the framework is trained on cleaner ground-truth *i.e.* AVA-Captions (CS or CWS) instead of AVA-Original (NS).

that our caption filtering strategy reduces the number of useless captions from the data and thus the network learns more accurate and informative components.

Generalizability

We wanted to test whether the knowledge gained by training on a large-scale but weakly annotated dataset is generic *i.e.* transferable to other image distributions. To do so, we train our frameworks on AVA-Captions and compare them with the models from [10], trained on PCCD. Everything is tested on the PCCD validation set. The models used by [10] are: (a) CNN-LSTM-WD is the NeuralTalk2 framework trained on PCCD. (b) Aspect oriented (AO) and (c) Aspect fusion (AF) are supervised methods, trained on PCCD. Note, that all the models are based on the NeuralTalk2 framework [147] and hence comparable in terms of architecture.

In Table 4.1(b), we observe that both CS and CWS outperform CNN-LSTM-WD and AO in S-1 scores. AF is still the best strategy for the PCCD dataset. Please note, both AO and AF are supervised strategies and require well defined “aspects” for training the network. Hence, as also pointed out in [10], it is not possible to train these frameworks on AVA as such aspect-level annotations are unavailable. However, we observe that both CS and proposed CWS, trained on AVA-Captions score reasonably well on PCCD. They are also generic strategies which can be easily mapped to other captioning tasks with weak supervision. We argue that the observed generalization capacity is due to training with a large and diverse dataset.

Subjective (Human) Evaluation

Human judgement is still the touchstone for evaluating image captioning, and all the previously mentioned metrics are evaluated based on how well they correlate with the same. Therefore, we perform quality assessment of the generated captions by a subjective study. Our experimental procedure is similar to Chang *et al.* [10]. We found 15 participants with varying degree of expertise in photography (4 experts and 11 non-experts) to evaluate our framework. In order to familiarize the participants with the grading process, a brief training with 20 trials was provided beforehand.



Figure 4.6: **Subjective evaluation of caption filtering:** The matrix compares our scoring strategy and human judgement for distinguishing a *good* and a *bad* caption. The rows stand for our output, and the columns represent what humans thought. We observe that the proposed caption filtering strategy is fairly consistent with what humans think about the informativeness of a caption.

The subjective evaluation was intended to assess: (a) whether the caption scoring strategy (Equation 4.2) is consistent with human judgement regarding the same (b) the effect of cleaning on the quality of generated captions.

(a) Consistency of Scoring Strategy: We chose 25 random images from the validation set, and from each image, we selected 2 accepted and 2 discarded captions. During the experiment, the subject was shown an image and a caption, and was asked to give a score on a continuous scale between 0 and 100. As an initial training, the subjects were presented with an equal number of accepted and discarded captions based on the proposed scoring strategy. They were not made aware of underlying working principle of the formula they were supposed to judge but were provided with a general understanding of what was meant by “informativeness” of a caption.

In Figure 4.6a and 4.6b, we plot our predictions and human judgement in a confusion matrix. We find that our strategy is fairly consistent with what humans think as a good or a bad caption. Interestingly, with the experts,

Category	κ_α	PLCC	SRCC
Expert	0.50	0.48	0.52
Non-Expert	0.55	0.55	0.61

Table 4.2: We measure inter-rater agreement for the scoring strategy using Krippendorff’s alpha. A value between 0 and 1 indicates positive agreement and therefore we find that our strategy is judged by human subjects quite reliably with $\alpha \geq .5$. On the other hand, correlation between the algorithm and human judgement regarding a caption is measured using PLCC and SRCC. We notice that the algorithm proposed is fairly consistent with the human judgement in both the metrics.

our strategy produces more false positives for bad captions. This is probably due to the fact that our strategy scores long captions higher, which may not always be the case and is a limitation. Note, that since our strategy was to make a binary choice between accepting or discarding a caption, we threshold the user input to create binary labels as accepted (good) or discarded (bad) and then compare using a confusion matrix.

In Table 4.2, we report the inter-rater agreement between the scores provided by different subjects using Krippendorff’s alpha. It gives a measure of how well the different subjects agree on the comment quality. We also report Spearman and Pearson correlation coefficients between the user input and the algorithm. This gives a measure of how well the proposed strategy correlates with human judgement.

(b) Effect of Caption Filtering: Similarly, 25 random images were chosen from the validation set. Each image had 3 captions, the candidates generated by NS, CS and CWS frameworks. During each trial, the subject was shown an image and one of the captions and asked to rate it into one of the categories - Good, Common and Bad. These discrete categories follow from [10] and the participants were asked to rate a caption based on whether it conveyed enough information about a photograph. Note, that similar to the previous experiment, the subjects were not told about the working principle of the underlying algorithms but only given a general description of a good, common or bad caption. We observe in Table 4.3 the percentage of good, common and bad captions generated by each method.

Method	Experts				Non-Experts			
	Good (3)	Com (2)	Bad (1)	Avg	Good (3)	Com (2)	Bad (1)	Avg
NS	0	80	20	1.80	0	84	16	1.84
CS	8	84	8	2.0	28	68	4	2.24
CWS	4	80	16	1.88	20	72	8	2.12

Table 4.3: **Subjective comparison of baselines:** We observe that human subjects find CS and CWS to be comparable but both significantly better than NS. This underpins the hypothesis derived from the quantitative results that filtering improves the quality of generated captions and the weakly supervised features are comparable with the ImageNet trained features

We observe that humans did not find any caption from NS to be good. Most of them were common or bad. This is due to its high tendency to generate the short, safe and common captions. Humans find CS to be performing slightly better than CWS which can probably be attributed to the lack of supervision during training the CNN. But as mentioned in Section 4.1, semi-supervised training is effective in practical scenarios due the easy availability of data and it might be worth investigating whether it is possible to improve its performance using more data and more complex representations. Additional qualitative results are provided in Figure 4.1.

4.6 Conclusion

In this work, we studied aesthetic image captioning which is a variant of natural image captioning. The task is challenging not only due to its inherent subjective nature but also due to the absence of a suitable dataset. To this end, we propose a strategy to clean the weakly annotated data easily available from the web and compile AVA-Captions, the first large-scale dataset for aesthetic image captioning. Also, we propose a new weakly-supervised approach to train the CNN. We validated the proposed framework thoroughly, using automatic metrics and subjective studies.

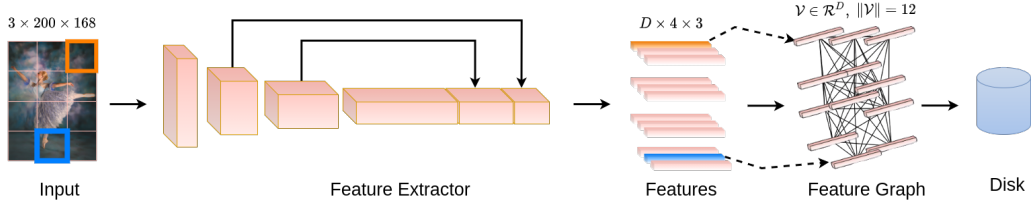
As future work, it could be interesting to explore alternatives for utilizing

the weak-labels and exploring other weakly-supervised strategies for extracting rich aesthetic attributes from AVA. It could also be interesting to extend this generic approach to other forms of captioning such as visual storytelling [91] or stylized captioning [90] by utilizing the easily available and weakly labelled data from the web.

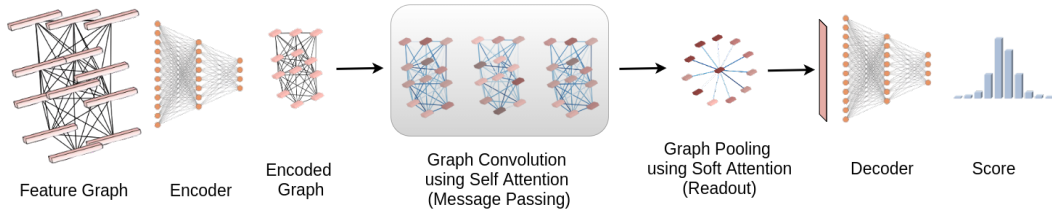
Chapter 5

Aspect Ratio and Layout Aware Aesthetic Score Regression

Aspect ratio and spatial layout are two of the principal factors influencing the aesthetic value of a photograph. However, incorporating these into the traditional convolution-based frameworks for the task of Image Aesthetics assessment is problematic. The aspect ratio of the photographs gets distorted while they are resized/cropped to a fixed dimension to facilitate training batch sampling. On the other hand, the convolutional filters process information locally and are limited in their ability to model the global spatial layout of a photograph. In this work, we present a two-stage framework based on graph neural networks and address both these problems jointly. First, we propose a feature-graph representation in which the input image is modelled as a graph, maintaining its original aspect ratio and resolution. Second, we propose a graph neural network architecture that takes this feature-graph and captures the semantic relationship between different regions of the input image using visual attention. Our experiments show that the proposed framework advances the state-of-the-art results in aesthetic score regression on the Aesthetic Visual Analysis (AVA) benchmark.



(a) **Feature Graph Construction:** In the first stage of the proposed pipeline, features ($D \times W \times H$) are extracted from the input image using a CNN, pre-trained on ImageNet. Instead of pooling to a fixed dimension, we split the features along D and construct a graph having a set of ($W \times H$) nodes, $\mathcal{V} \in \mathcal{R}^D$. A certain node in the feature graph corresponds to a certain neighbourhood in the input image (colour coded) and thus this representation preserves the spatial layout of the input. On the other hand, the structure and size of the graph captures the aspect ratio and resolution of the original image (e.g. for an input size $3 \times 200 \times 168$, the downsampled feature size is $D \times 4 \times 3$ and therefore the graph has 12 nodes).



(b) **Score regression using Graph Attention Network:** In the second stage, the high-dimensional input graph is first encoded to a compact representation and then passed to the GNN which performs global relational reasoning using three message passing and a pooling layer. The output is passed to a decoder which maps the features to the score. More details on the architecture are provided in Table.5.1.

Figure 5.1: **The proposed two stage pipeline**

5.1 Motivation

Image Aesthetics Assessment (IAA) refers to the task of predicting a score or rating of an image based on its aesthetic value. In photography, the aesthetic value of an image is influenced by several independent and correlated factors. For example, apart from the content, a photograph may look pleasing or dull for the photographer’s choice of colour balance, exposure levels, sharpness, content layout, crop etc. In the past, non deep learning or feature-based attempts tried to identify and encode these factors and define a generic model for Image Aesthetics [21, 22, 23, 24, 25, 26].

Because of the ambiguous and overlapping nature of aesthetic properties, the task was quite complex and ill-posed. But in recent years, deep learning based data-driven approaches have proven quite effective [7, 1, 2, 27, 28]. Due to the availability of the large scale Aesthetic Visual Analysis (AVA) [29] dataset and the rapid improvements in CNN architectures, the current state-of-the-art methods outperform the classical approaches by a wide margin. However, most deep learning-based methods face two primary challenges — aspect ratio awareness and image layout understanding.

Aspect ratio *i.e.* the ratio of the height and width of the photograph plays a crucial role in Image Aesthetics. For example a portrait shot or *selfie* may look better in a 1 : 1 ratio but a 16 : 9 frame may be better suited for a wide landscape or architectural shot. But in a standard CNN set up, the inputs have to be scaled or cropped to a uniform size in order to facilitate mini-batch sampling. This pre-processing step results in distortion or loss of information and subsequently introduces an undesirable bias in the learned features. Lu *et al.* [2] proposed a solution by sampling multiple crops from the input and aggregating them and thereby roughly approximating the entire image. Several other multi-crop based strategies have followed [55, 151]. The problem with these multi-crop approaches is computational cost and also their sensitivity towards the crop sampling and aggregation strategy. A different approach is to use a dynamic receptive field for the input and use an adaptive or average pooling layer at a later stage in the network [28, 7]. While this approach allows to use an arbitrary sized input in its original resolution, the pooling operators nevertheless map it to a fixed size at the feature-level and thus discard the aspect ratio.

The second challenge *i.e.* image layout understanding refers to capturing the spatial relations of the important visual elements of a photograph. The placement of the different objects within a frame is a key factor in photographic composition and there are several principles such as symmetry, The Rule of Thirds, framing *etc.* that photographers exploit to add aesthetic value to their images. Standard CNNs by design, have local receptive fields to achieve translation-invariance and are limited in their ability to perform global relational reasoning [152]. Towards this Lu *et al.* [1]

proposed a two column local-global approach in which the local column processed crops and the global column processed a warped version of the entire image. But it lacked aspect ratio understanding as discussed before. Ma *et al.* [27] used an object detector to detect salient regions in the image and then fused the coordinates with the main network. But apart from computational overhead due to the subnet, the number of objects for every image was experimentally set to five, which is not true for all images.

In this work, we address both these problems jointly, using graph neural networks (GNN) [100]. GNNs have two key advantages over CNNs. First, they are designed to work for arbitrary sized graphs. Secondly, they are able to capture both local and non-local dependencies between nodes efficiently, using a technique called neural message passing [103]. Based on this, we propose a strategy in which arbitrary sized images are represented as arbitrary graphs and then we develop a framework leveraging the power of GNNs and address the two problems discussed above. Specifically, there are two distinct stages in our pipeline — (a) **Feature Graph Construction** (Figure 5.1a) (b) **Score Regression Using GNN** (Figure 5.1b). (a) First, an ImageNet-trained CNN-based feature extractor computes features from an input image in its original aspect ratio and resolution. The features across different layers are concatenated along the depth dimension thereby capturing both the high and low level details from the image. Next, a feature graph is constructed from the concatenated maps such that the aspect ratio and the spatial relations of the input image are encoded in the structure of the graph. For example, as can be seen in Fig 5.1a, for an input with feature dimensions $(D \times 4 \times 3)$, the resulting feature graph is a set of 12 nodes, $\mathcal{V} \in \mathbb{R}^D$. The position of a node v_i corresponds to a certain region in the image. (b) In the next stage, a mini-batch of graphs is sampled from the disk and fed to a GNN based framework for aesthetic score regression. The architecture consists of an encoder, an attention based graph-convolution block and a decoder. The encoder maps the high dimensional sparse graph to a compact representation. The encoded feature-graph is fed to the graph convolution layers which performs spatial reasoning using message passing and pooling. Finally, it is passed to a decoder for predicting the scores.

Apart from maintaining the aspect-ratio and layout information, the proposed feature-graph representation has other advantages. Aesthetic quality is influenced by factors such as high frequency components (texture, sharpness etc.) and the low frequency elements (patterns, shapes etc.). On one hand, while cropping the image preserves the original high frequencies, a part of the low frequency information such as patterns and shapes is lost. On the other hand, resizing the entire input does a better job at preserving the ‘global’ components but distorts and blurs the image [1]. We avoid this typical ‘*catch-22*’ situation by extracting features from the entire input in its original resolution. Moreover, by concatenating features from the different layers of the CNN we are able to encode diverse information across the frequency spectrum [7]. We do not pool or resize the CNN features to a fixed dimension at any stage of the pipeline and thus the feature graph representation encodes rich visual information, aspect ratio and layout, simultaneously.

On their part, GNNs too have several benefits. They can be trained efficiently with mini-batches of arbitrarily sized graphs *i.e.* each element in the batch can have a different structure. Traditional convolutional frameworks followed by fully connected layers (CNN-FC) require the samples of a batch to have the same dimensions and thus lack this advantage. Additionally, graph convolutions are capable of capturing long range dependencies unlike the traditional CNNs which have local receptive fields. Thus, the proposed GNN block efficiently utilizes unique properties stored in each feature-graph and learns robust features for the target task. In this work, instead of using the traditional message passing [100] for graph convolutions, we base our framework on a variant that uses multi-headed self-attention [105], where nodes are combined selectively based on the image layout and content. We discuss more about this in Sec.5.2.

Essentially, we utilize the rich representational power of CNNs for modelling appearance while exploiting GNNs for a better semantic understanding. We evaluate our idea on the AVA dataset, which is the largest publicly available widely used benchmark for Image Aesthetics and advance the state-of-the-art results for aesthetics score regression. The summary of our contributions is as follows:

Contributions

1. We propose a novel graph-based, aspect-ratio aware representation for CNN features extracted from images in their original resolution.
2. We propose a GNN based framework using visual attention which is both aspect ratio preserving and layout aware.
3. We advance the state-of-the-art results for aesthetics score regression on the AVA dataset.

This chapter is organized as follows. In Sec. 5.2 we discuss our contributions in detail, in Sec. 5.3 we discuss the evaluation metrics and baselines used and report our results.

5.2 Pipeline

In this section we present the details of the proposed pipeline. First we elaborate on our two main contributions. In Sec.5.2.1, we discuss the proposed feature-graph construction and in Sec.5.2.2, we present the theory and architectural details of our GNN block and in Sec.5.2.3 we state the training procedure.

5.2.1 Feature Graph Construction

The first stage of our pipeline can be roughly divided into two sub-stages: feature extraction and graph construction.

Feature Extraction: We chose Inception-Resnet [74] architecture as the backbone network for extracting robust visual features. The choice is motivated from previous works [56, 7] which have demonstrated that Inception networks perform better for regression as compared to other popular choices such as ResNet [5] or DenseNet [153]. Following [7], we use pre-trained ImageNet [49] weights directly and did not notice any significant difference from fine-tuning the backbone on AVA dataset. This is probably due to the fact that although they are tuned for object recognition, these weights capture generic visual properties as also observed in several other tasks such as segmentation [120], style-transfer [121], captioning

[32] etc.

Generally, in an Inception block [154], input feature maps are handled by different filter sizes (e.g. $1 \times 1, 3 \times 3, 5 \times 5$), in parallel and then concatenated before passing it to the next layer. The multiple receptive fields let the network process the input at different scales. Inception-Resnet consists of 43 such blocks with residual connections. We collect the feature maps after each inception block, resize them to match the spatial resolution of the final layer feature maps and concatenate. For example as shown in Fig 5.1a, with an input of size 200×168 , the last layer features are downsampled to a size of 4×3 and feature maps from all the previous 43 inception blocks are resized accordingly and concatenated resulting in a feature size of $16928 \times 4 \times 3$. As discussed in Sec 5.1, extracting *multi-level* features helps in capturing a wider range of image frequencies and has been tried before for Image Aesthetics in [7, 54, 3].

Graph Construction: Once the multi-level features are extracted from an input, constructing the feature-graph $G(V, E)$ is straightforward, where V, E represents the set of nodes and edges, respectively. The feature map $F(I) \in \mathbb{R}^{D \times W \times H}$ is split along D i.e. the depth dimension into a set of node vectors $V \in \mathbb{R}^D$ as shown in Figure 5.1a. Note, that for Inception-Resnet, $D = 16928$ and $\|V\| = W * H$ e.g. 12 for Fig 5.1a. We construct a complete graph i.e. each node is connected with an undirected edge to every other node without self loops and therefore $\|E\| = \binom{W*H}{2}$. Stating formally,

$$G(V, E) \leftarrow \bigoplus_{i=1}^L \mathcal{T}[f_i(I)] \quad (5.1)$$

where, \bigoplus denotes concatenation, L stands for the number of layers in the feature extractor, $f_i(I) \in \mathbb{R}^{d_i \times w_i \times h_i}$ represents feature at layer i and \mathcal{T} is a resize operation.

The proposed feature-graph representation has three important properties. First, unlike the traditional CNN-FC pipelines, we avoid pooling the features from all images to a fixed size. Due to this, the number of nodes in the graph is a function of W and H and hence proportional to the input resolution and aspect ratio, which is unique to each image. Second, the

spatial layout of the input is preserved as shown in Fig 5.1a, where a certain region of the input (coloured boxes) corresponds to a specific node in the graph. Third, by constructing a complete graph, we ensure that the long range dependencies of the input are captured by graph convolution layers in the next stage. The problem with CNN-FC frameworks while capturing long range or ‘global’ dependencies has recently gained significant attention and is an active area of research [152, 155, 156]. The issue is particularly important in the case of Image Aesthetics as photographic composition often involves ‘globally’ aligning subjects within the frame in a certain way such as The Rule of Thirds.

The feature-graphs are constructed separately for all the images of AVA dataset and stored in the disk as HDF5 files in 16-bit floating point numpy arrays. While training the GNN block in the next stage, the feature graphs are loaded as mini-batches directly from the disk.

5.2.2 Score Regression using GNN

In this section, we discuss the second stage of our pipeline where the input is a feature-graph and the output is the aesthetic score distribution as illustrated in Fig 5.1b. Graph-regression tasks such as ours can broadly be divided into two stages [103] — (a) **Message Passing**: Each node updates itself by exchanging information within its neighbourhood and the output, which is also a graph with the same structure, encodes the complex correlations between the different nodes. (b) **Readout**: The nodes of the encoded graph are combined using a function, which maps an arbitrary number of tensors to a fixed-sized vector, which is generally mapped to the desired output using a fully connected network. In the following sections, we describe the details of these two stages in our framework.

Message Passing with Self-Attention

Given an input graph $G(V, E)$, the traditional message passing [100] in GNNs is formally stated as follows:

$$v_i' = \sum_{j \in \mathcal{N}_i} \mathcal{W}v_j \quad (5.2)$$

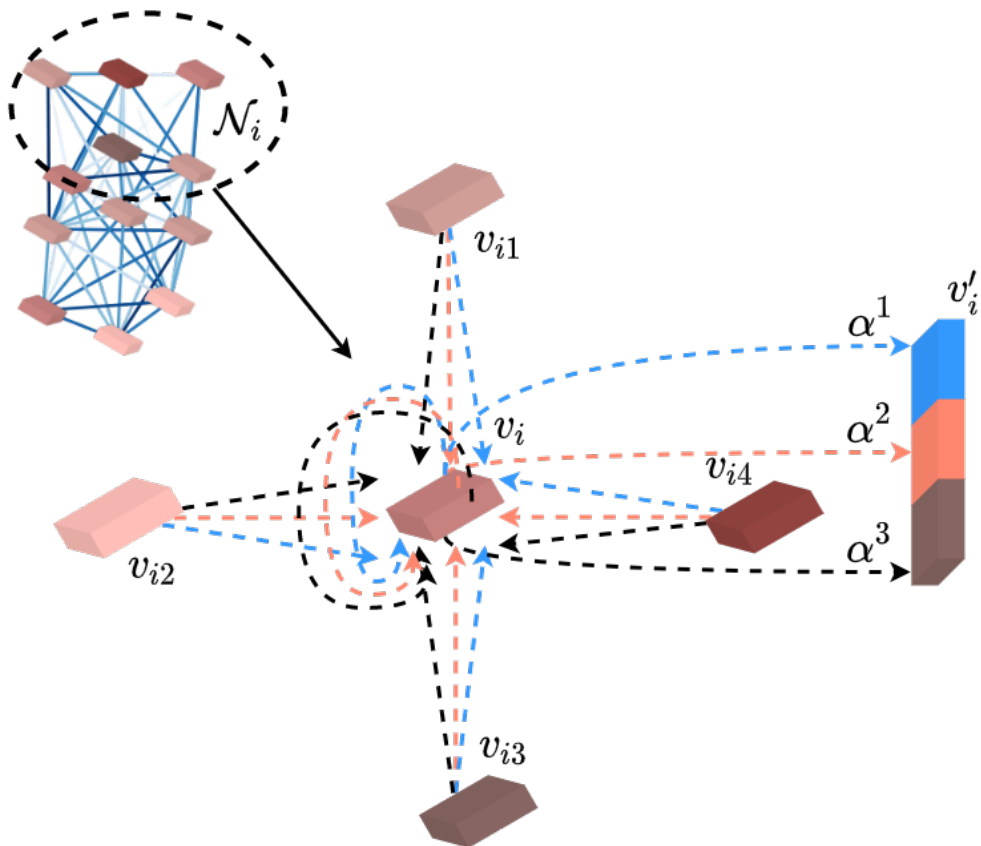


Figure 5.2: **Message Passing with Self-Attention:** A toy scenario for the update $v_i \rightarrow v'_i$, with four neighbours and three attention heads (red, blue and black). v'_i is the concatenation of the output from the different attention heads (Eq 5.6). Note, that this step is repeated for every node $v_i \in G$ and the output is also a graph with the same structure as the input.

where, \mathcal{N}_i is the neighbourhood of node v_i , v'_i is the updated node and \mathcal{W} is a shared learnable weight matrix. It is similar to traditional convolutions except there, \mathcal{N}_i is a fixed grid and its size is equal to the receptive field of the filter. Eq 5.2, can be formalized as a GPU trainable matrix operation $V' = \mathcal{W}\mathcal{A}V$, where \mathcal{A} is the adjacency matrix that stores the neighbourhood information. A problem with the traditional message passing is that while updating a node, it assigns equal weights to its neighbours. This is undesirable for Image Aesthetics since certain areas of the image may *drive* the composition more than the rest (such as the eyes in a portrait) and it is important that this relationship is efficiently encoded.

To this end, Veličković *et al.* [105] extended the traditional message passing algorithm using self-attention [157] and proposed graph attention networks (GAT). Eq 5.2, for GAT is modified to:

$$v'_i = \sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathcal{W}v_j \quad (5.3)$$

where the attention coefficient α_{ij} captures the effect of v_j on v_i . The form of attention used in this work follows [105] which can be formally stated as:

$$\alpha_{ij} = \frac{\exp\left(g(v_i, v_j)\right)}{\sum_{k \in \mathcal{N}} \exp\left(g(v_i, v_k)\right)} \quad (5.4)$$

Eq 5.4, is a soft-max over $g(v_i, v_j)$, where g is a neural network of the form:

$$g(v_i, v_j) = \text{LeakyReLU}\left(g(\mathcal{U}v_i \oplus \mathcal{U}v_j)\right) \quad (5.5)$$

where \mathcal{U} is a shared linear transformation, typically another neural network, applied to each node. In practice to increase stability during training, instead of a single α_{ij} , multiple attention heads are concatenated together and the final form of message passing (modified from Eq 5.3) used

is:

$$v'_i = \bigoplus_{k=1}^K \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathcal{W}^k v_j \right) \quad (5.6)$$

where K is the number of attention heads.

Fig. 5.2 illustrates Eq. 5.6 using a simple scenario with four neighbours and three attention heads. Unlike in Eq. 5.2 where a node treats all its neighbours equally, in Eq. 5.6, neighbours are *attended* selectively based on the weight α_{ij} , which is in fact a function of v_i and v_j . Essentially, modelling the input as feature graph preserves its aspect-ratio and resolution whereas using self-attention lets us encode the global correlations of the input effectively. Note, that the output of this layer which is passed to the next block, is an encoded feature-graph with the same structure as the input.

Readout with Soft-Attention

Our readout function is based on the soft-attention based global pooling layer [158] (GATP), which takes the arbitrarily-structured encoded graph and generates a fixed-sized embedding. It is formally stated as follows:

$$\mathcal{G}_{pool} = \sum_{n=1}^{||\mathcal{G}||} \text{softmax} \left(h_{\text{gate}}(v_n) \right) \odot v_n \quad (5.7)$$

where h_{gate} is a neural network that generates the attention mask and \mathcal{G}_{pool} is the pooled graph. We extend this to a multi-headed approach, as in the previous section as follows:

$$\mathcal{G}_{pool} = \frac{1}{K} \sum_{k=1}^K \sum_{n=1}^{||\mathcal{G}||} \text{softmax} \left(h_{\text{gate}}^k(v_n) \right) \odot v_n \quad (5.8)$$

where K different attention heads h_{gate}^k are learnt. Note, that unlike Eq.5.6, the output from the K attention heads are averaged instead of concatenation.

Block	Module
Encoder	Linear (16928,2048) ReLU() BatchNorm(2048)
Message Passing	Dropout (0.8) GAT($\mathcal{U} = \text{Linear} (2048, 64)$, K = 16) ReLU() GraphSizeNorm() ×3
Readout	GATP ($h_{gate} = (2048,1)$, K = 16)
Decoder	Dropout(0.8) Linear (2048,1024) ReLU() BatchNorm(1024) Linear (1024,10)

Table 5.1: Architectural Details

Architecture

The final architecture used for the task as illustrated in Fig. 5.1b consists of an encoder, the message passing layer, readout layer and a decoder. The architectural details are presented in Table 5.1. The encoder is a fully-connected layer followed by ReLU activation and batch-norm. The encoder maps the high-dimensional feature-graph to a more compact latent representation which is fed to the graph layers. The message passing layer in the graph block consists of three graph attention layers [105] with the shared linear transformation $\mathcal{U} = \text{Linear} (2048,64)$ and 16 attention heads. Each layer is preceded by a dropout regularizer and followed by a ReLU activation and GraphSizeNorm layer [159]. The GraphSizeNorm Layer normalizes the node features by the graph size *i.e.* $\mathcal{G}_{norm} = \frac{\mathcal{G}}{\|\mathcal{G}\|}$. This is followed by the readout or pooling layer with 16 attention heads and a fully connected encoder $h_{gate} = \text{Linear} (2048,1)$. Finally, a decoder takes the output of the pooling layer and maps it to the score distribution. It consists of two fully connected layers with a ReLU activation, batch-norm, preceded by a dropout. We implemented the entire framework using PyTorch [160] and PyTorch-Geometric [161].

5.2.3 Training

We train the network using the traditional mean-squared-error loss between the normalized histogram of scores (1×10) and the network output. The choice of the loss function is motivated from [7]. We tried other commonly used distribution-matching loss function such as Earth Mover Distance [56], but did not notice any significant improvement in performance. We use the ADAM optimizer [162] with default PyTorch parameters. The starting learning rate is set to $1e-4$ which is reduced every epoch by a factor $(1 - \frac{e}{E})^\lambda$, where e and E are the current and the total number of epochs, respectively. λ and E is experimentally determined as 2.5 and 30, respectively. We followed an augmentation strategy similar to [7] to add regularization. Four corner crops each covering 85% of the image and with the original aspect ratio were extracted and flipped giving eight different representations. During training, one random augmentation was chosen and during inference the scores were averaged. Using a batch size of 64 on a Nvidia RTX 2080-Ti 11 GB GPU, training a model until convergence takes about 9 hours.

5.3 Experiments

5.3.1 Dataset and Metric

We evaluate our approach using the AVA dataset, which is a collection of 230,000 train and 20,000 test images. The images were uploaded by photographers for competitions hosted on www.dpchallenge.com and rated by the community on a scale of 10. The ground-truth annotations for AVA are provided as a 10-bin histogram of scores. The final score is obtained as a weighted average of the histogram. With these scores, there are two traditional tracks for evaluation.

One is to pose the problem as a classification task by labelling the images as “good” or “bad” based on a cut off score and subsequently measure classification accuracy. But this approach is problematic for several reasons [53, 56, 7]: First, the choice of the threshold (*typically* set to 5) is quite arbitrary as the average rating for AVA dataset is 5.5 and the performance has been found to be highly sensitive to slight variations of this thresh-

old. Second, AVA is highly unbalanced with 7 : 3 ratio for good and bad samples. A biased model predicting good and bad samples with 100% and 50% (*i.e.* random) precision, respectively, could achieve an accuracy of 85%, significantly higher than any state-of-the-art method. A *balanced* or weighted accuracy score is thus a better measure for AVA, which unfortunately is reported by only a handful of previous methods. Third, a 0/1 labelling scheme fails to capture the relative aesthetic ranks of photographs, a feature desirable in many real world applications for multimedia content creation.

The second and more robust evaluation strategy is adopted by methods which pose the problem as a regression task and predict the scores directly and measure the Pearson (PLCC) and Spearman (SRCC) Rank Correlation Coefficients between the predicted and ground-truth scores. They are widely applied for Image Quality Assessment (IQA) and are better suited for capturing the entire range of scores while avoiding arbitrary thresholds and label imbalance. Hence, we chose to optimize our framework for the score regression task and used PLCC and SRCC for the ablation study and comparison with the current state-of-the-art [56, 7, 6]. Nevertheless, an indirect measure of accuracy and balanced accuracy (\mathcal{T}_{Acc} and $\mathcal{T}_{Acc(B)}$) was computed by thresholding the output at 5 and our framework was also compared with the classification-based approaches [29, 2, 163, 53, 28, 27, 57] for a holistic understanding of the performance.

5.3.2 Ablation study

Here, we investigate the effects of the different components of our pipeline. Specifically, we study the effects of the encoder-decoder and the benefits of using attention in the graph layers over the conventional message passing and readout layers. For that, we define the following baselines:

- (a) **Avg-Pool-FC**: This is the most basic network where the Inc-ResNet features are average pooled and trained using a single fully-connected (FC) layer (16928×1).
- (b) **Avg-Pool-ED**: The FC layer from Avg-Pool-FC is replaced by the encoder-decoder blocks from Table 5.1.

Method	PLCC (μ/σ)	SRCC (μ/σ)	\mathcal{T}_{Acc}	$\mathcal{T}_{Acc(B)}$
Avg-Pool-FC	0.712/ 0.13	0.710/ 0.13	80.21	70.63
Avg-Pool-ED	0.744/ 0.31	0.741/ 0.30	81.17	72.50
GCN-GMP	0.759/ 0.32	0.757/ 0.31	81.77	73.38
GAT _{x1} -GMP	0.761/ 0.33	0.759/ 0.32	81.82	74.46
GAT _{x1} -GATP	0.762/ 0.33	0.760/ 0.31	81.83	74.58
GAT _{x3} -GATP	0.764/ 0.35	0.762/ 0.34	82.15	76.32

Table 5.2: **Ablation Study:** We start with the most basic single fully connected layer (Avg-Pool-FC) and gradually add the different components namely, the encoder-decoder, feature graph, message passing and readout. We notice steady improvements in the performance in all metrics.

- (c) **GCN-GMP:** Instead of average pooling, the feature-graph representation is introduced. We use the traditional message passing [100] without attention and global mean pooling or averaging for readout. The encoder-decoder is identical as the previous baseline.
- (d) **GAT_{x1}-GMP:** We replace the traditional message passing of GCN-GMP with Eq.5.6 (1 layer).
- (e) **GAT_{x1}-GATP:** The readout function from GAT_{x1}-GMP is replaced by Eq.5.8.
- (f) **GAT_{x3}-GATP:** We add 2 more layers of message passing to GAT_{x1}-GATP. This is our final framework with all the different components discussed in Sec.5.2.

In Tab.5.2, we compare the performance of these different baselines in terms of PLCC and SRCC scores between the mean score (μ) and standard deviation of the score distribution (σ). While a good correlation between the mean scores is important and obvious, the standard deviation measures how well the true and predicted distributions are aligned. This is crucial especially for the borderline images, which depend on a finer and nuanced judgment. We notice that Avg-Pool-FC has the lowest scores, which is expected as it preserves neither the aspect ratio nor the layout. Avg-Pool-ED performs slightly better, probably because the encoder-decoder layers provide additional non-linearity. The performance improves in GCN-

GMP, where the proposed feature-graph representation is introduced. Note, that even this simple graph baseline performs better than MLSP (Pool-3FC)[7], the current state-of-the-art for score regression (in Tab. 5.3). The superiority of GCN-GMP over the pooling-based strategies underpins the hypothesis that aspect-ratio and layout information is crucial for IAA which is otherwise lost due to pooling or resizing. On the other hand in $\text{GAT}_{\times 1}$ -GMP, attention based message passing utilizes this information more efficiently by focusing on important regions. The performance improves more in $\text{GAT}_{\times 1}$ -GATP by replacing the readout block with the soft-attention based pooling. Finally, we add two more layers to the message passing block and this completes the proposed framework $\text{GAT}_{\times 3}$ -GATP.

As we added more layers, we noticed overfitting and to handle that, we introduced dropout regularization before each new block. We also noticed that the performance was quite sensitive to the parameters of the normalization layers (both graph and batch normalization). The number of attention heads and the encoded graph size also mattered. All these hyper-parameters were determined experimentally and we urge the reviewers to check the supplementary material or the code for more details. In Figure 5.3, we plot the score distribution of the different baselines, averaged over all the test samples and plot them with the ground-truth distribution. A significant difference is observed between the best (f) and worst (a), with marginal improvements between the rest in between. This is consistent with the quantitative results in Table 5.2.

5.3.3 Comparison with the State-of-the-Art

In Table 5.3, we compare our approach with the previous score-regression benchmarks and report PLCC, SRCC and \mathcal{T}_{Acc} . We chose three different baselines from NIMA [56]. The baselines use different feature-extractors. We chose two different architectures proposed in MLSP [7]. We also select [6], which is to the best of our knowledge, the most recent work on regression-based IAA. We notice that our method outperforms the rest in terms of all the metrics. It can also be observed that both [7] and our method outperform [56, 6] by a large margin. This is probably due to the fact that both methods extract features at their original resolution whereas the others use some mode of cropping or warping to achieve a

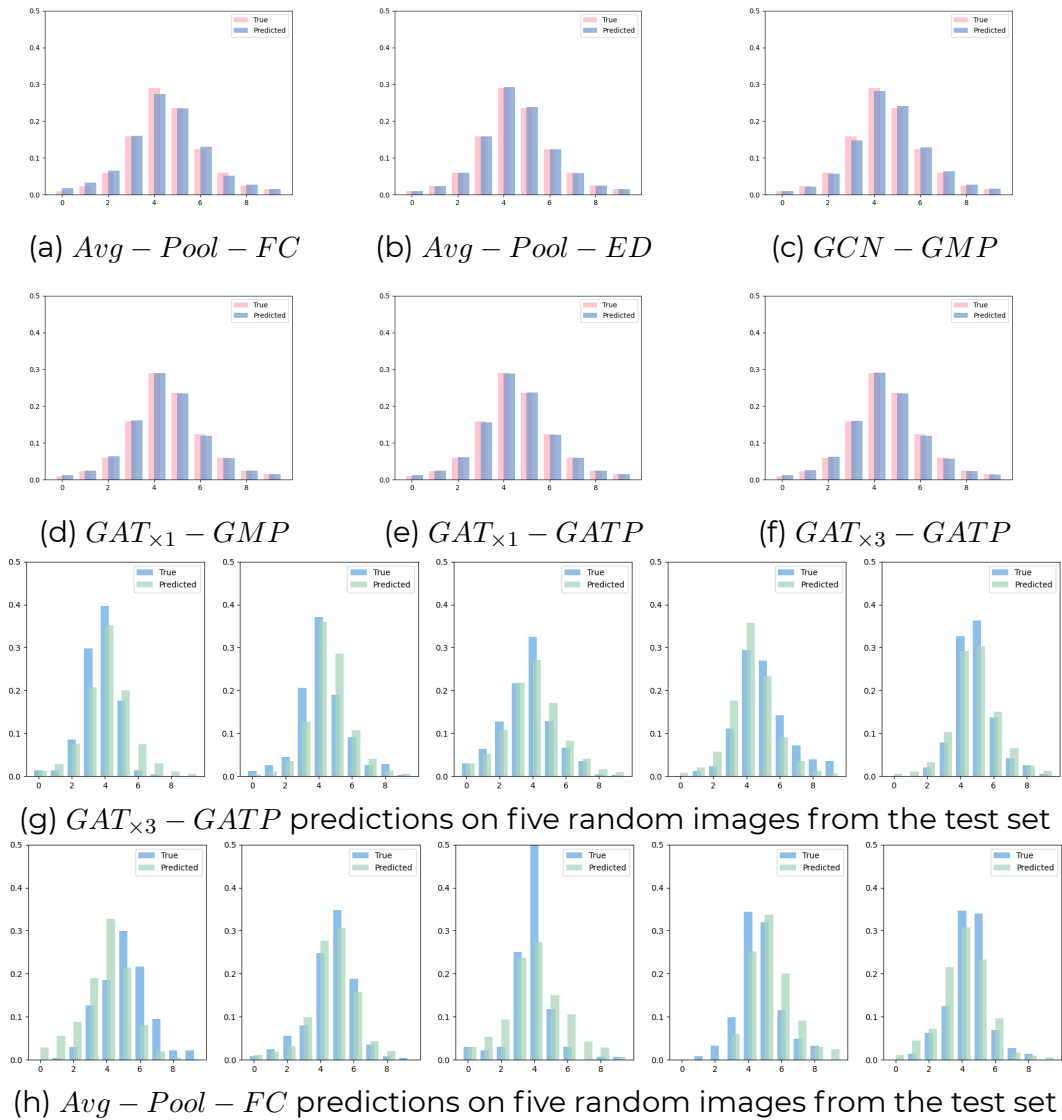


Figure 5.3: (a)-(f) Average score distribution of different baselines plotted with the ground truth distribution. (g)-(h) Baseline predictions on five random images from the test set

Method	PLCC	SRCC	\mathcal{T}_{Acc} %
NIMA (Mob Net) [56]	0.518	0.510	80.36
NIMA (VGG 16) [56]	0.610	0.592	80.60
NIMA (Inc V2) [56]	0.636	0.612	81.51
Attn-based Spatial [6]	0.710	0.707	80.48
MLSP (Single-3FC) [7]	0.745	0.743	81.37
MLSP (Pool-3FC) [7]	0.757	0.756	81.72
GAT_{x3}-GATP	0.764	0.762	82.15

Table 5.3: **PLCC, SRCC, \mathcal{T}_{Acc}** : Our approach outperforms the previous methods for score regression in all the metrics. To the best of our knowledge, [6] is the most recent work on this topic and [7] is the state-of-the-art.

uniform input size. Our results underpin the claim that seeing the image in their original resolution benefits, which is also pointed out by [7]. On the other hand, we perform better than [7]. Since we use the same Inception-Resnet based feature extractor, we argue that the improvement is primarily due to the aspect-ratio aware representation and a better layout understanding by virtue of using graph networks.

In Table 5.4, we compare with previous methods which approach this problem differently *i.e.* as a classification task. Of the baselines selected, AVA [29], MNA-CNN-Scene [28], A-Lamp [27]¹, MP_{ada} [55] and RGNet [57] report traditional accuracy only but DMA-Net-ImgFu [2], New Rapid [163] and DAN-2 [53] report both traditional and balanced accuracy. DAN-2 uses two different sampling strategies to handle the label imbalance in AVA and we selected both. We observe that the proposed approach performs better than all the baselines except [55, 57] in terms of traditional accuracy. But, note that ours is \mathcal{T}_{Acc} , indirectly computed from the scores whereas their networks are optimized for binary classification loss directly. As pointed out in [7], it is not ideal to judge the performance of a network optimized for regression using accuracy. For example, in Table 5.3, the three different NIMA baselines differ by a large margin in terms of correlation scores but only slightly in terms of \mathcal{T}_{Acc} . This is probably due to the fact that a better correlation score essentially means a better under-

¹We compared with the same A-LAMP baseline as [7] that uses no auxiliary information.

Method	Acc %	Acc (B) %
AVA [139]	67.0	-
DMA-Net-ImgFu [2]	75.4	62.80
New RAPID [163]	75.42	61.77
DAN-2 (Balanced Sampling) [53]	75.96	73.51
MNA-CNN-Scene [28]	77.4	-
DAN-2 (Regular Sampling) [53]	78.72	69.45
A-Lamp [27]	81.70	-
MP _{ada} [55]	83.03	-
RGNet [57]	83.59	-
GAT_{x3}-GATP	82.15	76.32

Table 5.4: **Accuracy (Acc) and Balanced Accuracy (Acc (B))**: We compare our regression-based approach using indirect thresholded accuracy (\mathcal{T}_{Acc}) with methods which pose the problem as a classification problem and are optimized with a binary classification loss. We find the performance comparable and better in terms of Acc and Acc (B), respectively.

standing of the score distribution and aesthetic ordering of images. As the correlation scores improve, the network probably learns more complex and nuanced aspects which distinguish different images, especially the ambiguous ones *i.e.* ones with similar scores. Such improvements are not reflected in the overall accuracy as the ordering of images does not matter when they belong to the same class.

Moreover as pointed out earlier, due to the label imbalance, accuracy alone may not reflect the true performance of a network. A network which performs poorly for the smaller category may well achieve very high overall numbers. A balanced or weighted accuracy is a better measure in this scenario. We observe that our approach performs better than the other baselines which report balanced accuracy. Our approach is in fact better than DAN-2 [53] which not only trains for classification but uses a balanced sampling strategy to handle the label imbalance explicitly.

5.3.4 Label Imbalance

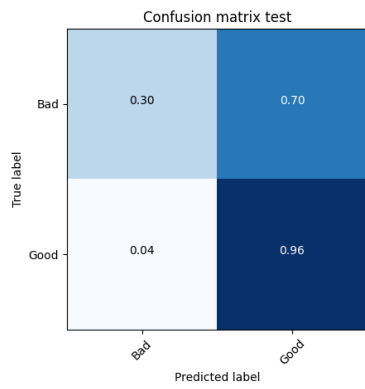
We plot the confusion matrices of three different baselines to investigate the label imbalance problem in AVA in Fig 5.4. (a) is the Avg-Pool-FC net-

work, trained with binary cross entropy loss for 0/1 classification. Note, the numbers are true precision computed from predicted labels (b) is the confusion matrix of Avg-Pool-FC trained for regression using MSE loss. Here, per class precision is computed from \mathcal{T}_{Acc} (c) is the confusion matrix of the proposed GAT_{x3} -GATP framework, trained using MSE loss for regression. We notice a considerable bias in (a) towards label 1 *i.e.* good images. This is due to the 7 : 3 label distribution in AVA. This bias is reduced significantly when the same network is trained for regression using MSE loss. This is probably because the network learns relative aesthetic ranks of images and subsequently more nuanced aspects of aesthetics. Such rank information is especially important for the borderline images *i.e.* score close to 5, most of which get miss-classified as 1 in the case of binary classification. However in (c), we notice a significant improvement in the performance of the 'bad' or sparse category. Clearly, the graph networks and the feature graph representation performs much better for the borderline images.

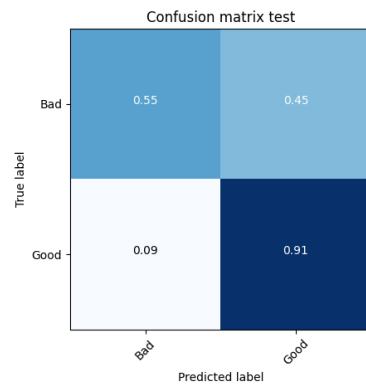
5.3.5 Qualitative Results

The images are displayed in their original aspect ratio in Figure 5.5. **Row 1** shows images with $GT \geq 6$ *i.e.* highly rated. The images are sharp, colourful and well exposed. We observe, that all the predictions are also ≥ 6 . This is indicative of the fact that the proposed network captures the appearance based characteristics quite well. **Row 2** consists of poorly rated images where $GT \leq 4$. These images have trivial appearance problems such as dull colours (first), blown out exposure (second and fifth), bad focus (third) Likewise, the network predictions are consistent with ground truth scores. **Row 3** displays the images with average rating with $4 \leq GT \leq 6$. These are quite challenging to handle. Many of these images such as the second and third one from left, are pleasing due to the inherent story or moment. Apart from appearance, the essence of the image can be understood from a host of other factors such as subject placement, motion capture, framing *etc.* We notice that the predicted scores are consistent with the GT here as well. We argue that this is due to the efficient handling of aspect ratio and layout.

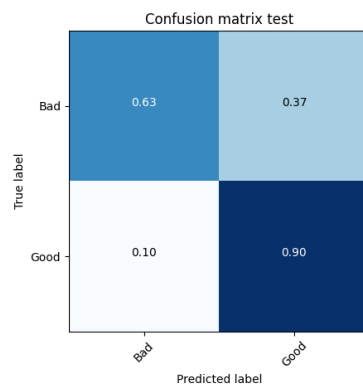
In summary, the proposed idea performs better than the previous state-



(a) Avg-Pool-FC + BCE



(b) Avg-Pool-FC + MSE



(c) $GAT_{\times 3}$ -GATP + MSE

Figure 5.4: **Confusion Matrices**

of-the-art for aesthetic score regression in terms of PLCC, SRCC, \mathcal{T}_{Acc} [7] and $\mathcal{T}_{Acc(B)}$ [53]. Our experiments suggest that the improvement is largely due to the rich information stored in the feature graph and its efficient utilization by the $GAT_{\times 3}$ -GATP framework.

5.4 Conclusion

In this work, we address two central challenges in deep-learning based Image Aesthetics assessment: aspect ratio and spatial layout understanding. We do that jointly, by combining the complementary representational powers of CNNs and GNNs, enhanced by visual attention. Our experiments verify that the proposed approach advances the state-of-the-art for the score regression task, significantly. In the future, there can be several possible improvements. For example, given the complexity of the task, it may be interesting to explore more advanced graph architectures. The framework may also benefit from exploring more options for the CNN backbone with better global-relational reasoning abilities. We hope that the method and results in this work will be useful for future research in Image Aesthetics and other related areas.

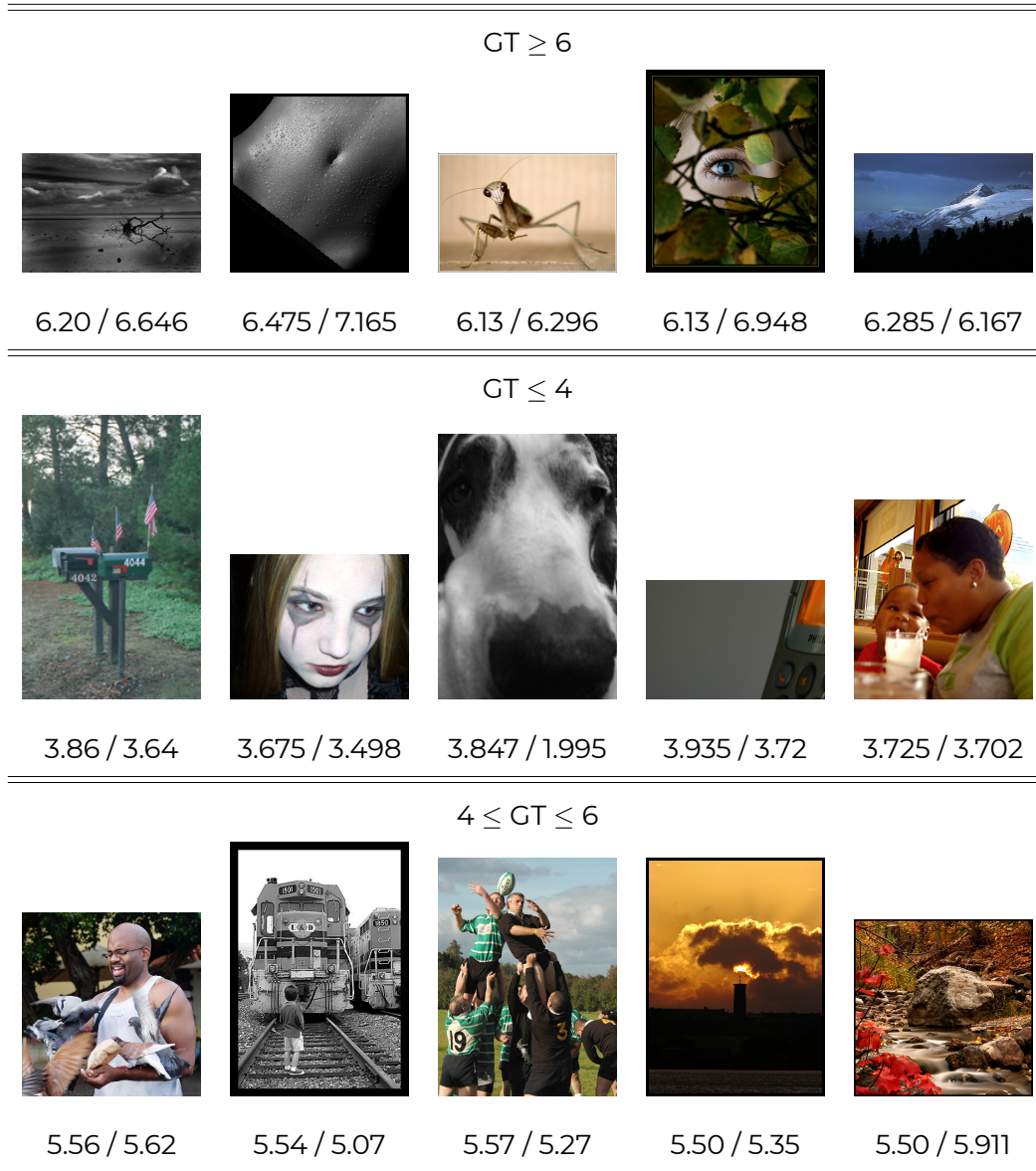


Figure 5.5: GAT_{x3} -GATP predictions / Ground truth (GT) scores for images randomly sampled from AVA:

Chapter 6

Conclusion

This brings us to the end of the dissertation. We studied different aspects of Image Aesthetics assessment and proposed deep learning based solutions which were validated by experimental studies. In this chapter we conclude the thesis by summarizing the contributions followed by revisiting the research question and discussing some of the potential research directions.

6.1 Summary

In Chapter 1, we discussed the factors that make the task of analyzing Image Aesthetics quite difficult even for humans — its subjective nature and the fact that it requires reasonable background knowledge in photography. In this context, we explored deep learning algorithms and studied the complexities of automating this process. Drawing motivation from the recent advances of deep learning for standard computer vision tasks, we proposed solutions to some of the problems associated with Image Aesthetics assessment.

In Chapter 2, we reviewed the classical and recent advances of Image Aesthetics assessment. We discussed AVA dataset, the largest and widely used benchmark for the task and the factors that make it challenging to work with. We also briefly highlighted the state-of-the-art CNN architectures and other related areas that motivated this research.

In Chapter 3, we analyzed the ability of deep learning based methods to learn aesthetic attributes like the Rule of Thirds, depth of field, vanishing-lines etc. We addressed the limitations of standard CNNs in understanding global layout owing to their *translation invariance* property. We introduced a novel input representation which is *geometry sensitive*, *position cognizant* and *appearance invariant*; and a two-column CNN architecture that performs better than the state-of-the-art in aesthetic attribute prediction.

In Chapter 4, we studied aesthetic image captioning. We proposed an automatic cleaning strategy of noisy web data to create a dataset ‘AVA-Captions’, ($\sim 230,000$ images with ~ 5 captions per image). Additionally, by exploiting the latent associations between aesthetic attributes, we proposed a strategy for training the CNN based visual feature extractor.

In Chapter 5, we explored graph neural networks for aesthetic score regression. We showed that the aspect ratio and global layout of a photographic image can be efficiently captured using a visual attention based graph neural network. Our experiments showed that the proposed framework advanced the state-of-the-art results in aesthetic score regression on the Aesthetic Visual Analysis (AVA) benchmark by a significant mar-



Figure 6.1: Correlated Attributes: Each row displays eight samples from two categories (four each) which were mutually confused. Row 1 is horror / noir and Row 2 is minimal / geometry

gin.

6.2 Outlook and Future Work

Research Question Revisited

Broadly, we tried to investigate — ***“How efficiently can artificial agents be trained for Image Aesthetic Analysis?”***.

In this context, we studied three different applications:

- Aesthetic Attribute Prediction.
- Feedback or Aesthetic Image Captioning.
- Aesthetic Score Prediction.

How do we feel about the overall performance of deep learning for the applications explored?

Our key findings and results indeed look promising. For example, in Chapter 3 we found that our network learnt the correlation between exposure settings and the overall emotion conveyed in a photograph (Figure 6.1, Row 1). Without direct supervision, it learnt that “horror” is probably related to dark exposures and high contrast, which is true. Similarly in row 2, we see confusions between minimal and geometry. But are these wrong inferences? The pictures having ground truth label as “minimal” could



i like the composition , but i think it would have been better if you could have gotten a little more of the building

this is a great shot . i love the way the light is coming from the left.

Figure 6.2: Realistic Comments

very well be tagged as "geometric" because the two attributes are quite overlapping. We found that neural networks learn far more than they are explicitly trained for and they develop a "generic" understanding of aesthetic attributes. Similarly, in Chapter 4, we noticed that in some cases, the network learnt several interesting factors and generated sensible quality feedback almost as good as a real human critic. For example, in Figure 6.2 the network understands that the left image is "tightly" cropped and the composition would benefit from adding some more context around the door. For the right image, it predicts the position of the light source and indeed lighting is the key element in that composition. We believe the fact that the network learns such subtle aspects of composition without any explicit guidance is remarkable.

How does it compare with a human critic?

For certain genres such as scenery or low light photography, which are "naturally" quite appealing we observed that the network does a decent job, both in terms of identifying the attributes and predicting the overall aesthetic value. Especially, when the salient attributes are based purely on appearance such as colour, contrast, sharpness etc., the network does reasonably well. However, when the key factors driving a composition are somewhat abstract such as "the decisive moment" or "a slice of life", the networks struggle. This is not unexpected, as these factors transcend the actual content of the images and refer to subjective feelings or larger so-

ciopolitical contexts. We as human beings gather such knowledge continuously from domains quite independent of photography and then as photographers apply it to capture or analyze a photograph. This is difficult for a machine, if not impossible using the standard principles of supervised learning.

Where does it go from here?

On one hand, in comparison to the traditional problems of vision such as object detection, segmentation, action recognition *etc.*, Image Aesthetics is a less explored domain. While there exists previous research in categorization/ score-regression and attribute prediction, our work on textual feedback is one of its first. The complexities involved in the tasks which were discussed in the previous chapters warrant further exploration.

On the other hand in deep learning, novel architectures are being proposed and newer challenges are being addressed on a regular basis. It became popular only about a decade ago and has taken over many areas such as computer vision, natural language processing, speech signal processing, *etc.* Nevertheless, as the amount of data continues to grow with the growth of internet applications in every sphere of our personal lives and broader scientific explorations, the scope of research in deep learning will continue to expand and diversify.

Immediate Goals: An important next step could be to curate a clean and multipurpose benchmark for Image Aesthetics. We discussed the problems of the AVA dataset in Section 2.2.2. The new dataset should ideally have clean and evenly distributed ground truth annotations, raw files, multi-labelled samples and enough data to train deep architectures. It could also be interesting to try out the latest architectural developments in deep learning such as non-local convolutional blocks [152], transformers [164] and more advanced graph-based networks. There exists a reasonable volume of research in natural scene parsing using semantic graphs [165]. Image aesthetics could potentially benefit from those approaches as well. The quality of textual feedback could be improved in terms of quality and diversity using the recent *reward-guided* strategies [166] for image captioning. Another track could be exploring generative models

for aesthetic quality enhancement, guided by aesthetic feedback.

Long-term Goals: In our opinion, the most critical and less explored aspect of Image Aesthetics is a well defined evaluation framework. As discussed in Chapter 5, the binary categorization of photographs is a very coarse measure of quality. Unfortunately, due to the lack of a well defined metric this has become the standard over the years. Moreover, metrics such as accuracy, precision scores measure the overall performance quantitatively, but do not correlate well with the subtleties of human judgments [56]. Another area could be to explore aesthetics driven generative models for image/video editing. For example, a system capable of generating aesthetically pleasing scene templates based on screenplays could benefit the animation industry. Image aesthetics could also help improving the quality of emerging multimedia content using virtual and augmented reality. Last but not the least, it could be interesting to explore evolutionary and continuous learning strategies by distilling knowledge from diverse domains to complement the abstract and ever-changing notion of “aesthetics”.

This research was a small step towards understanding a complex problem and we hope it proves somewhat useful for future explorations.

Appendix A

Abbreviations

Short Term	Expanded Term
AVA	Aesthetic Visual Analysis
AIC	Aesthetic Image Captioning
CNN	Convolutional Neural Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
LDA	Latent Dirichlet Allocation
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
LSTM	Long Short-Term Memory
MAP	Mean Average Precision
MSE	Mean Squared Error
NIC	Natural Image Captioning
PCP	Per Class Precision
PCCD	Photo Critique Captioning Dataset
PLCC	Pearson Correlation Coefficient
RNN	Recurrent Neural Network
RGB	Red Green Blue
RoT	Rule of Thirds
SRCC	Spearman Correlation Coefficient

Appendix B

Aesthetic Attributes

1. **Complementary Colours :** This refers to the colours on the opposite side of the colour wheel. (See Figure B.1). In a photograph, it is often desirable to have elements complementing each other in terms of colour

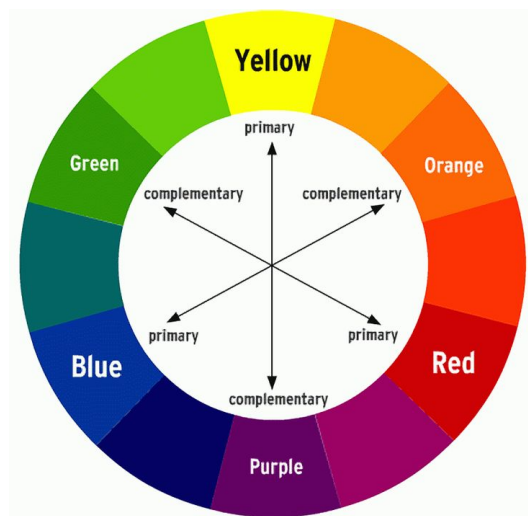


Figure B.1: Colour Wheel : Example of complementary colours are red-green, yellow-purple combinations.

2. **Duotones :** It refers to images having two-tonal ranges. Usually, one dark colour is used as the base and two or more are used as secondary colours for the highlights and shadows. Duotones are differ-

ent from grayscale images.

3. **HDR** : A photograph which has a high dynamic range comes under this category. In other words, it refers to a nicely exposed photograph, in which the difference between the brightest and darkest regions is high. It is crucial in HDR that there is some detail in both the highlights and shadows.
4. **Image Grain** : It refers to the presence of noise in the picture. In the past, pictures used to be grainy because of the limited capability of the camera sensor. Nowadays, often noise is intentionally introduced in an image to give it a retro appearance.
5. **Light on White** : This category refers to light-coloured objects captured in a white background. This style is typically used in close-up shots to direct the viewer to the main foreground object with least background interference.
6. **Long Exposure** : These kind of pictures are captured by keeping the shutter opened for a longer duration to allow more light to come in. It is a technique often used to perform light-painting or astrophotography.
7. **Macro** : This refers to the shots of tiny objects using a very narrow depth-of-field. It is very common while photographing food, flowers or insects.
8. **Motion Blur** : This is used in scenarios to capture motion in a picture. With a moderately low shutter speed a high movement is usually blurred. It might not be desirable in some scenarios but photographers use it to incorporate a sense of motion.
9. **Negative Image** : It is just an inverted representation of the luminance of the image. The whites are black and the blacks are white. It is an artistic effect, often used to create surreal effects to photographs.
10. **Rule of Thirds (RoT)** : This rule is concerned with the alignment of the main subjects within a photograph. It states that if an image is aligned with a 3×3 grid, then it is more aesthetically pleasing to put the subjects in certain locations. See Figure B.2. Thus in portrait

photography, the eyes are usually aligned along that line. In macro, it is the object in focus, which is positioned at one of those points. Thus, RoT is a purely geometry based property which can be combined with any other photographic property for a good composition.

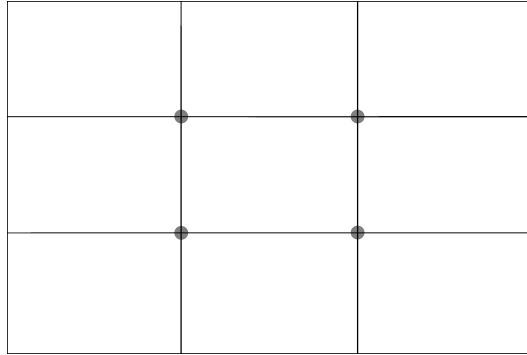


Figure B.2: Rule of Thirds : It is observed that when the main subject is placed at one of the four points instead of the centre of the photograph, it is aesthetically more pleasing.

11. **Shallow Depth of Field :** This style is close to macro style of photography but the objects are not necessarily small. It involves a wide aperture, resulting in a very shallow depth of field. As a result, the main subjects are in focus, whereas, the background is blurred.
12. **Silhouettes :** These photographs are usually captured with the camera facing the light source. As a consequence, the the objects which are between the camera and the light are dark, as in a solar eclipse. This creates a nice shadow or outline of the object without any other details inside. This style is often used to take landscape shots
13. **Soft Focus :** It is a technique which is applied by using lenses that cause spherical aberration. In other words it blurs the objects in the image while retaining the strong edges. It is often used to soften the rough areas in the photograph. For example one uses this technique in portrait photography to smoothen the skin to wash away bruises or dark spots..
14. **Vanishing Point :** Also termed as *leading-lines*, it involves the use of

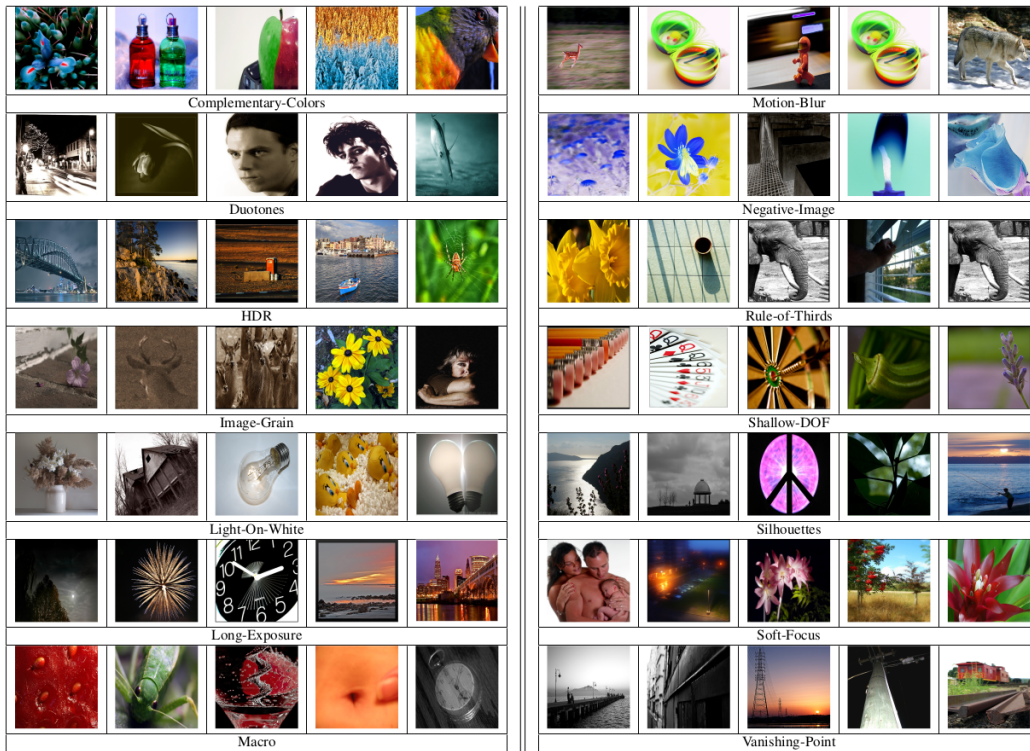


Figure B.3: Some photographic attributes from AVA and their descriptions

natural lines in the scene (for example railway tracks, horizon, etc.), to lead the viewer towards the main subject. It is a widely used compositional attribute, especially applied in landscape photography.

Bibliography

- [1] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 457–466, ACM, 2014.
- [2] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pp. 990–998, 2015.
- [3] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell, "Recognizing image style," in *BMVC 2014*, 2014.
- [4] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [6] Y. Xu, Y. Wang, H. Zhang, and Y. Jiang, "Spatial attentive image aesthetic assessment," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2020.
- [7] V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9375–9383, 2019.
- [8] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting hu-

- man eye fixations via an lstm-based saliency attentive model," *arXiv preprint arXiv:1611.09571*, 2016.
- [9] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5561–5570, 2018.
- [10] K.-Y. Chang, K.-H. Lu, and C.-S. Chen, "Aesthetic critiques generation for photos," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 3534–3543, IEEE, 2017.
- [11] B. Barnbaum, *The Art of Photography: A Personal Approach to Artistic Expression*. Rocky Nook, Inc., 2017.
- [12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [14] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.
- [15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.
- [16] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1520–1528, 2015.
- [17] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.
- [18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

- [19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150–1157, IEEE, 1999.
- [20] A. Hertzmann, "Can computers create art?," in *Arts*, vol. 7, p. 18, Multidisciplinary Digital Publishing Institute, 2018.
- [21] R. Datta, D. Joshi, J. Li, and J. Wang, "Studying aesthetics in photographic images using a computational approach," *Computer Vision–ECCV 2006*, pp. 288–301, 2006.
- [22] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, pp. 419–426, IEEE, 2006.
- [23] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," *Computer Vision–ECCV 2008*, pp. 386–399, 2008.
- [24] P. Obrador, M. A. Saad, P. Suryanarayan, and N. Oliver, "Towards category-based aesthetic models of photographs," in *International Conference on Multimedia Modeling*, pp. 63–76, Springer, 2012.
- [25] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1657–1664, IEEE, 2011.
- [26] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.
- [27] S. Ma, J. Liu, and C. Wen Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [28] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 497–506, 2016.

- [29] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2408–2415, IEEE, 2012.
- [30] T. Barrett, *Criticizing photographs*. McGraw-Hill, 2005.
- [31] K. Ghosal, M. Prasad, and A. Smolic, "A geometry-sensitive approach for photographic style classification," *arXiv preprint arXiv:1909.01040*, 2019.
- [32] K. Ghosal, A. Rana, and A. Smolic, "Aesthetic image captioning from weakly-labelled photographs," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.
- [33] X. Zheng, T. Chalasani, K. Ghosal, S. Lutz, and A. Smolic, "Stada: Style transfer as data augmentation," *arXiv preprint arXiv:1909.01056*, 2019.
- [34] O. Yadav, K. Ghosal, S. Lutz, and A. Smolic, "Frequency-domain loss function for deep exposure correction of dark images," *Signal, Image and Video Processing*, pp. 1–8, 2021.
- [35] N. Rosenblum, *A world history of photography*. Abbeville Press New York, 1997.
- [36] A. Scharf, *Art and photography*. Penguin books New York, 1983.
- [37] "ALBRIGHT-KNOX ART GALLERY international exhibition of pictorial photography." <https://www.albrightknox.org/>, Retrieved January 10, 2018.
- [38] N. Carroll, *Philosophy of art: a contemporary introduction*. Routledge, 2012.
- [39] F. H. Goodyear, "Between amateur & aesthete: The legitimization of photography as art in america 1880-1900," *American Studies International*, vol. 39, no. 2, pp. 102–102, 2001.
- [40] K. Lubben, *Magnum contact sheets*. Thames & Hudson New York, 2011.
- [41] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan, et al.,

- “Light field photography with a hand-held plenoptic camera,” *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.
- [42] S. Peleg, Y. Pritch, and M. Ben-Ezra, “Cameras for stereo panoramic imaging,” in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1, pp. 208–214, IEEE, 2000.
- [43] W. Luo, X. Wang, and X. Tang, “Content-based photo quality assessment,” in *2011 International Conference on Computer Vision*, pp. 2206–2213, IEEE, 2011.
- [44] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, “Assessing the aesthetic quality of photographs using generic image descriptors,” in *2011 international conference on computer vision*, pp. 1784–1791, IEEE, 2011.
- [45] T. O. Aydın, A. Smolic, and M. Gross, “Automated aesthetic analysis of photographic images,” *IEEE transactions on visualization and computer graphics*, vol. 21, no. 1, pp. 31–42, 2015.
- [46] C. Li and T. Chen, “Aesthetic visual quality assessment of paintings,” *IEEE Journal of selected topics in Signal Processing*, vol. 3, no. 2, pp. 236–252, 2009.
- [47] C. Li, A. C. Loui, and T. Chen, “Towards aesthetics: A photo quality assessment and photo selection system,” in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 827–830, 2010.
- [48] L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, and J. Li, “Oscar: On-site composition and aesthetics feedback through exemplars for photographers,” *International Journal of Computer Vision*, vol. 96, no. 3, pp. 353–383, 2012.
- [49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, IEEE, 2009.
- [50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,”

- in *European conference on computer vision*, pp. 740–755, Springer, 2014.
- [51] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, “Photo aesthetics ranking network with attributes and content adaptation,” in *European Conference on Computer Vision*, pp. 662–679, Springer, 2016.
- [52] G. Malu, R. S. Bapi, and B. Indurkha, “Learning photography aesthetics with deep cnns,” 2017.
- [53] Y. Deng, C. C. Loy, and X. Tang, “Image aesthetic assessment: An experimental survey,” *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.
- [54] Y.-L. Hii, J. See, M. Kairanbay, and L.-K. Wong, “Multigap: Multi-pooled inception network with text augmentation for aesthetic prediction of photographs,” in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1722–1726, IEEE, 2017.
- [55] 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*, pp. 879–886, 2018.
- [56] H. Talebi and P. Milanfar, “Nima: Neural image assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.
- [57] D. Liu, R. Puri, N. Kamath, and S. Bhattachary, “Composition-aware image aesthetics assessment,” *arXiv preprint arXiv:1907.10801*, 2019.
- [58] Q. Chen, W. Zhang, N. Zhou, P. Lei, Y. Xu, Y. Zheng, and J. Fan, “Adaptive fractional dilated convolution network for image aesthetics assessment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14114–14123, 2020.
- [59] “Adobe Photoshop.” <https://www.adobe.com/ie/products/photoshop.html>. Website Link.
- [60] “Snapseed.” https://play.google.com/store/apps/details?id=com.niksoftware.snapseed&hl=en_IE&gl=US. Website Link.

- [61] "Adobe Blog." <https://blog.adobe.com/en/publish/2020/10/20/photoshop-the-worlds-most-advanced-ai-application-for-creatives.html#gs.vkkyh0>. Website Link.
- [62] "Instagram." <https://www.instagram.com/>. Website Link.
- [63] "Facebook Inc." <https://www.facebook.com>. Website Link.
- [64] "Shutterstock Inc." <https://www.shutterstock.com/>. Website Link.
- [65] "EyeEm." <https://www.eyeem.com/>. Website Link.
- [66] "EveryPixel." <https://www.everypixel.com/>. Website Link.
- [67] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.
- [68] I. Rish et al., "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, pp. 41–46, 2001.
- [69] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [70] Y. LeCun, "The mnist database of handwritten digits," <http://yann.lecun.com/exdb/mnist/>, 1998.
- [71] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.
- [72] K. Chellapilla, S. Puri, and P. Simard, "High performance convolutional neural networks for document processing," in *Tenth International Workshop on Frontiers in Handwriting Recognition*, Suvisoft, 2006.
- [73] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for

- image classification,” in *Twenty-second international joint conference on artificial intelligence*, 2011.
- [74] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” *arXiv preprint arXiv:1602.07261*, 2016.
- [75] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, “Deep networks with stochastic depth,” in *European Conference on Computer Vision*, pp. 646–661, Springer, 2016.
- [76] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.
- [77] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, “Grounded compositional semantics for finding and describing images with sentences,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.
- [78] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every picture tells a story: Generating sentences from images,” in *European conference on computer vision*, pp. 15–29, Springer, 2010.
- [79] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2text: Describing images using 1 million captioned photographs,” in *Advances in neural information processing systems*, pp. 1143–1151, 2011.
- [80] Y. Jia, M. Salzmann, and T. Darrell, “Learning cross-modality similarity for multinomial data,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2407–2414, IEEE, 2011.
- [81] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*, pp. 2048–2057, 2015.
- [82] J. Johnson, A. Karpathy, and L. Fei-Fei, “Densecap: Fully convolutional localization networks for dense captioning,” in *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4565–4574, 2016.

- [83] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, *et al.*, “From captions to visual concepts and back,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1473–1482, 2015.
- [84] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.
- [85] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, “Generation and comprehension of unambiguous object descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.
- [86] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, “Deep compositional captioning: Describing novel object categories without paired training data,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–10, 2016.
- [87] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, “Deep captioning with multimodal recurrent neural networks (m-rnn),” *arXiv preprint arXiv:1412.6632*, 2014.
- [88] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, 2015.
- [89] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, “Learning like a child: Fast novel visual concept learning from sentence descriptions of images,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2533–2541, 2015.
- [90] A. Mathews, L. Xie, and X. He, “Semstyle: Learning to generate

- stylised image captions using unaligned text,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8591–8600, 2018.
- [91] R. Kiros, “neural-storyteller,” 2015.
- [92] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” *arXiv preprint arXiv:1411.2539*, 2014.
- [93] T. L. Berg, A. C. Berg, and J. Shih, “Automatic attribute discovery and characterization from noisy web data,” in *European Conference on Computer Vision*, pp. 663–676, Springer, 2010.
- [94] X. Chen, A. Shrivastava, and A. Gupta, “Neil: Extracting visual knowledge from web data,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1409–1416, 2013.
- [95] S. K. Divvala, A. Farhadi, and C. Guestrin, “Learning everything about anything: Webly-supervised visual concept learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3270–3277, 2014.
- [96] C. Sun, C. Gan, and R. Nevatia, “Automatic concept discovery from parallel text and visual corpora,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2596–2604, 2015.
- [97] S. Vittayakorn, T. Umeda, K. Murasaki, K. Sudo, T. Okatani, and K. Yamaguchi, “Automatic attribute discovery with neural activations,” in *European Conference on Computer Vision*, pp. 252–268, Springer, 2016.
- [98] T. Yashima, N. Okazaki, K. Inui, K. Yamaguchi, and T. Okatani, “Learning to describe e-commerce images from noisy online data,” in *Asian Conference on Computer Vision*, pp. 85–100, Springer, 2016.
- [99] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, “Spectral networks and locally connected networks on graphs,” *arXiv preprint arXiv:1312.6203*, 2013.

- [100] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.
- [101] M. Henaff, J. Bruna, and Y. LeCun, “Deep convolutional networks on graph-structured data,” *arXiv preprint arXiv:1506.05163*, 2015.
- [102] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in *Advances in neural information processing systems*, pp. 3844–3852, 2016.
- [103] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” *arXiv preprint arXiv:1704.01212*, 2017.
- [104] M. Fey, J. Eric Lenssen, F. Weichert, and H. Müller, “Splinecnn: Fast geometric deep learning with continuous b-spline kernels,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 869–877, 2018.
- [105] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” *arXiv preprint arXiv:1710.10903*, 2017.
- [106] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, “Convolutional networks on graphs for learning molecular fingerprints,” in *Advances in neural information processing systems*, pp. 2224–2232, 2015.
- [107] J. Atwood and D. Towsley, “Diffusion-convolutional neural networks,” in *Advances in neural information processing systems*, pp. 1993–2001, 2016.
- [108] W. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Advances in neural information processing systems*, pp. 1024–1034, 2017.
- [109] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” in *Advances in Neural Information Processing Systems*, pp. 3856–3866, 2017.
- [110] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and

- J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," *arXiv preprint arXiv:1807.03247*, 2018.
- [111] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, pp. 577–585, 2015.
- [112] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, pp. 694–711, Springer, 2016.
- [113] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.
- [114] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, pp. 2377–2385, 2015.
- [115] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, pp. 3320–3328, 2014.
- [116] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.
- [117] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 2556–2565, 2018.
- [118] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

- [119] A. Rana, P. Singh, G. Valenzise, F. Dufaux, N. Komodakis, and A. Smolic, "Deep tone mapping operator for high dynamic range images," *Transaction of Image Processing*, 2019.
- [120] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
- [121] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, 2016.
- [122] A. Kell, "Where does data come from?," 2018.
- [123] D. Gershgorn, "The data that transformed ai research—and possibly the world," 2017.
- [124] E. Zerman, A. Rana, and A. Smolic., "Colornet - estimating colorfulness in natural images," in *The International Conference on Image Processing (ICIP)*, 2019.
- [125] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060, 2017.
- [126] Z. Ren and Y. Jae Lee, "Cross-domain self-supervised multi-task feature learning using synthetic imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 762–771, 2018.
- [127] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [128] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318, Association for Computational Linguistics, 2002.
- [129] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the*

- IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.
- [130] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Spice: Semantic propositional image caption evaluation,” in *European Conference on Computer Vision*, pp. 382–398, Springer, 2016.
- [131] E. Loper and S. Bird, “Nltk: The natural language toolkit,” in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, (Stroudsburg, PA, USA), pp. 63–70, Association for Computational Linguistics, 2002.
- [132] T. Nagarajan and K. Grauman, “Attributes as operators: factorizing unseen attribute-object compositions,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 169–185, 2018.
- [133] I. Misra, A. Gupta, and M. Hebert, “From red wine to red tomato: Composition with context,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1792–1801, 2017.
- [134] R. Santa Cruz, B. Fernando, A. Cherian, and S. Gould, “Neural algebra of classifiers,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 729–737, IEEE, 2018.
- [135] M. A. Sadeghi and A. Farhadi, “Recognition using visual phrases,” in *CVPR 2011*, pp. 1745–1752, IEEE, 2011.
- [136] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, “Visual translation embedding network for visual relation detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5532–5540, 2017.
- [137] J. Ramos *et al.*, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 133–142, 2003.
- [138] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000.
- [139] L. Marchesotti, N. Murray, and F. Perronnin, “Discovering beautiful at-

- tributes for aesthetic image analysis," *International journal of computer vision*, vol. 113, no. 3, pp. 246–266, 2015.
- [140] D. Parikh and K. Grauman, "Relative attributes," in *2011 International Conference on Computer Vision*, pp. 503–510, IEEE, 2011.
- [141] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1637–1644, 2014.
- [142] J. Wang, Y. Cheng, and R. Schmidt Feris, "Walk and learn: Facial attribute representation learning from egocentric video and contextual data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2295–2304, 2016.
- [143] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang, "Learning hypergraph-regularized attribute predictors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409–417, 2015.
- [144] S. Li, "Topic modeling and latent dirichlet allocation (lda) in python," 2018.
- [145] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 569–577, ACM, 2008.
- [146] R. Y. Rubinstein and D. P. Kroese, *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media, 2013.
- [147] R. Luo, "An image captioning codebase in pytorch," 2017.
- [148] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic eval-*

uation measures for machine translation and/or summarization, pp. 65–72, 2005.

- [149] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” *Text Summarization Branches Out*, 2004.
- [150] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, “Learning to evaluate image captioning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5804–5812, 2018.
- [151] L. Wang, X. Wang, T. Yamasaki, and K. Aizawa, “Aspect-ratio-preserving multi-patch image aesthetics score prediction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.
- [152] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.
- [153] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- [154] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.
- [155] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “Gcnet: Non-local networks meet squeeze-excitation networks and beyond,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.
- [156] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, “Axial-deeplab: Stand-alone axial-attention for panoptic segmentation,” *arXiv preprint arXiv:2003.07853*, 2020.
- [157] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,

- Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- [158] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," *arXiv preprint arXiv:1511.05493*, 2015.
- [159] V. P. Dwivedi, C. K. Joshi, T. Laurent, Y. Bengio, and X. Bresson, "Benchmarking graph neural networks," *arXiv preprint arXiv:2003.00982*, 2020.
- [160] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," *openreview.net*, 2017.
- [161] M. Fey and J. E. Lenssen, "Fast graph representation learning with pytorch geometric," *arXiv preprint arXiv:1903.02428*, 2019.
- [162] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [163] 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.
- [164] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.
- [165] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1261–1270, 2017.
- [166] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7008–7024, 2017.