Division of Computing and Mathematics
Faculty of Natural Sciences
University of Stirling

Paying Attention to Image Aesthetic Quality Assessment

By

de Sigley, Frida

Student Number: 2935530

Dissertation submitted in partial fulfillment for the the degree of
Master of Science in Artificial Intelligence.

February 28, 2022

Declaration of own work

I understand the nature of plagiarism, and I am aware of the University's policy on this. I certify that this dissertation reports original work by me during my University project except for the following:

- Transformer code used in experiments chapter **??** in was taken from Facebook (now Meta) AI GitHub[1] and taken in modified form[2]
- Pre-Trained models used in chapter **??** courtesy of Ross Wightman[3]
- The training loop in was adapted from PyTorch example of training loop provide in a notebook
- Formative model training of various transformers Was made using code from Phil Wang[4]
- While all other code was written by the author, there are some sections where code has been adapted from other sources - such as loss function examples[5]
- Unless explicitly cited, all figures are originally produced by the author. Where figures are taken from publications, citation(s) are included in the image caption; where images are taken from data-sets and then manipulated graphically, the associated data-set of the publication is used. For instance, figure **??** is from MNIST**LeCun1998**, but the image plot, patches, and graphical manipulation are original.
- Reproducing results from **Sheng2018a**, **Hosu2019**, **Talebi2018**, **Ma2017** was done using associated GitHub repositories[6] [7] [8] [9]
- The dissertation was proof read Laura Whyte for typographic errors and miss spellings.

Frida de Sigley

February 28, 2022

[1] https://github.com/facebookresearch/convit
[2] https://github.com/mawady/convit
[3] https://github.com/rwightman/pytorch-image-models
[4] https://github.com/lucidrains/vit-pytorch
[5] https://github.com/clcarwin/focal_loss_pytorch
[6] https://github.com/Openning07/MPADA
[7] https://github.com/subpic/ava-mlsp
[8] https://github.com/idealo/image-quality-assessment
[9] https://github.com/GuillaumeBalezo/A-Lamp

Acknowledgements

I would like to thank Dr Mohamed Elsayed Elawady, Dr Deepayan Bhowmik, Dr Kevin Swingler for there ongoing patience with my incessant questioning; their sound academic guidance and their genuine moral support, because deep learning is a deep subject. I would also like to thank my partner Laura Whyte for ongoing support and proof reading and thank my parents Linda Nelson and Jeremy Penford for their support throughout. Finally I would like to thank the Marian Dunbar for being such and enthusiastic proponent of Scottish Data and Tech and the DataLab Scotland whose funding made undertaking an MSc in Artificial Intelligence possible.

Abstract

This thesis examines Image Aesthetic Quality Assessment(IAQA) as a computer vision problem. We compare deep Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) and Convolutional Vision Transformers (ConViTs), and examine which models best predict 'image quality' as a binary classification problem.

While CNNS have superseded hand-crafted approaches to feature extraction, achieving this has required hand-crafting attention mechanisms as high level training policies on very complex CNN architectures - which are, themselves, time consuming and inefficient to train. We therefore produce a side-by-side comparison to assess whether self-attention can perform well as an IAQA classifier.

We perform training on the AVA Bench-marking dataset and show that in many cases both ViTs and ConVits outperform CNNs in side-by-side comparisons. Further, while ViTs and ConVits both require lengthy pre-training on very large datasets, they are excellent candidates for domain adaptation - often with pre-trained models performing well when architectural adjustments are made to output layers.

Surprisingly, this requires fewer training epochs than pre-trained CNN models to adapt to new domains. Further, while CNNs quickly overfit on the AVA Training subset, this is not the case with transformers. We also show that the conditions that suit each type of network differ, however, ConVits do not appear to require as many 'warm-up' epochs when being trained using transfer learning as they do in when being initially trained.

We show that while models will train on the AVA benchmarking dataset without pre-training, that using non pre-trained models does not achieve high training accuracy.

Total Words: 15046 Headers: 83 Math Inline: 126 Math Display: 17

Total Words	15046
Headers	83
Inline Equations	126
Equations	17

Contents

List of Tables

Listings

List of Figures

1 Introduction

1.1 Background

> We are of course, supposing for the present that the questions are of the kind to which an answer 'Yes' or 'No' is appropriate, rather than questions such as 'What do you think of Picasso?'**TURING1950**
>
> *Alan M. Turing*

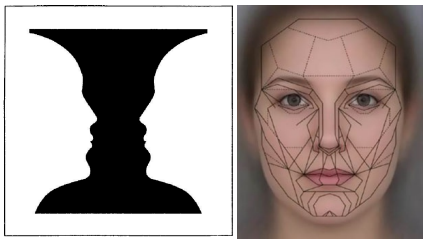
sal2021, **Raiyani2021**, **Helber2017**, **Mazzia2020**, **Baamonde2019** Image aesthetic assessment as a computational and machine learning task is important, challenging and fascinating. The application of aesthetics is wide, and encompasses areas such as cosmetics, calligraphy, and painting alongside beauty prediction of digital portraits**Redi2015a** and facial beauty prediction**Gray2010**, **Eisenthal2006a**, **Zhang2016**. Applications within the art domain include style recognition**Cho2020**, **Fernando2021** and aesthetic visual quality of handwriting**Sun2015**. All of these approaches require some form of digital imagery, and this thesis will focus on the Aesthetic Quality Assessment of digital images in the form of digital photographs.

Within computer vision, there are sub-domains where aesthetics is a component such as Image Quality Assessment(IQA)**Gu2020**, **Seshadrinathan2009**, **Sheikh2005**, **Ke2006**, **Chubarau2021**, **Reisenhofer2018** and Image Restoration, which have used metrics such as PSNR and SSIM. However, Image Aesthetic Quality Assessment(IAQA) treats aesthetic prediction as a target rather than focusing necessarily on individual image properties. Where IQA and image restoration have focused on restoring objective image distortions and are aimed at modelling or predicting a physical property such that it can be removed, IAQA focuses on predicting classes or aesthetics attributes.

While aesthetic is often based on individual subjective perception, it is important to demarcate this from objective and empirical study within the field of machine learning and pattern recognition. Where the focus is on objective ground truth, whether by experts or aggregated scores, the computer vision task is to best predict ground truth. The perceptual element, alongside the potential for ambiguity and lack of consensus on concepts such 'beauty' and 'aesthetic appreciation'**Datta2006**, add to the complexity and the challenge of IAQA**Yang2019** as a machine learning problem. One aspect of this is that it includes human visual perception, which has been shown within psychological studies to be both contingent upon the target photograph's context**Mullennix2020**, where the order in which images are viewed effects aesthetic perception, and to differ significantly between subjects**Younes2016**.

IAQA's reliance upon high-level social concepts**Datta2006** is however part of what makes the field so fascinating; with the potential to further insights into human visual attention, perception, and cognition[1] and continue a valuable tradition of biologically inspired computing**Hassabis2017** that began in the 1950s, in addition of course to IAQA's significant commercial and practical applications.

Computer vision has made significant gains in tasks such as object recognition, classification, medical imaging**Esteva2021**. The problem of 'vision' is however far from solved; for example, semantic scene detection remains a challenging area of focus. This is complex in part because it requires consideration of whether image segmentation *or* recognition comes first**Kreiman2021** (a task that can also be a challenge for humans) (see figure 1.1). IAQA is one such challenging domain application.



(a) Vase Face Figure Ground ambiguity f **Hobbs1990**, **Eysenk2012**

(b) An average face with a Phi mask not fitting the average face well **Zhang2016**

Figure 1.1: Examples of Ambiguity for Machines and Humans

Traditional approaches to IQA have aimed at measuring image distortion by degrees**gonzalez2008digital** within the wider field of digital image restoration. IQA's focus on modelling the non-commutative and non associative process of restoring digital image distortions caused by the corpuscular nature of light such as diffraction, refraction or image sensor noise caused during the capture of digital images in low light conditions.**szeliski2011computer**, **gonzalez2008digital** have informed approaches to early approaches IAQA. In both IAQA and IQA, traditional computer vision has been superseded by Machine Learning(ML) in many areas, using techniques such as Deep Learning using convolutional neural networks (CNNs) which have been successfully employed since first used for character recognition**LeCun198** **LeCun1998**.

This thesis will examine how these more recent techniques have been applied to challenging domain application of IAQA, and seek to improve and demonstrate effectiveness of different deep learning approaches within IAQA.

[1] *Vision* it might be argued it is a form of perception, property of cognition and intelligence**russell2016artificial** . Pigeons for instance have been shown to be remarkable observers of pathology **Levenson2015** when rewarded appropriately.

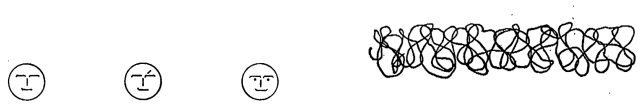
1.1.1 The Application Domain

The Oxford English Dictionary(OED) defined 'aesthetics' as 'Of or relating to the perception, appreciation, or criticism of that which is beautiful'**OED2021**. Aesthetics more widely includes appreciation of all senses, including touch and olfactory sensation**Hayes2015**, and further cognitive activities such as mathematical problem solving which have an element of reward or pleasure, where we might consider a solution 'elegant' or 'beautiful'**Rolls2014**. Aesthetics, therefore, can be considered a broad application domain that requires narrowing.

Attempts to study aesthetics have been made within psychology (imperial aesthetics)**Greb2017**, art criticism, philosophy, and as the subject of enquiry within the field of computer vision. IAQA, therefore, is a subset at the intersection of computer vision, aesthetics and image processing, focusing on digital photographic images. This might be considered a sub-domain within the wider field of visual aesthetics, where quality estimation or ranking is involved, and includes painting, calligraphy, cartoon imagery and sculptures. This section will provide a high-level overview of these interrelated fields and provide the framework for formulation IAQA as a computer vision problem.

1.1.2 Aesthetics Philosophy

Philosophy gives the the earliest examples of formal enquiry into aesthetics by Plato**Plochmann1976** (230-bc) and (1790) in the west by E. Kant and E. Burke **Kant1892kant**, **Burke1773philosophical**. This enquiry has continued into the 20th century with philosophers such as L. Wittgenstein**Wittgenstien1** who, in making general observations on aesthetic, remarks that enumerable facial expressions can be highlighted by only subtle changes in four strokes (figure 1.2a) contrasting it with figure 1.2b. Wittgenstein remarks that such squiggles drawn one after the other would be indistinguishable; a humorous, insightful remark that demonstrates clearly how aesthetic consideration is both highly nuanced and complex. One might ask here if this is the case for machines?



(a) Faces **(b)** Squiggle S

Figure 1.2: Left Faces, Right 'Squiggle S' **Wittgenstien1967**

1.1.3 Empirical Aesthetics

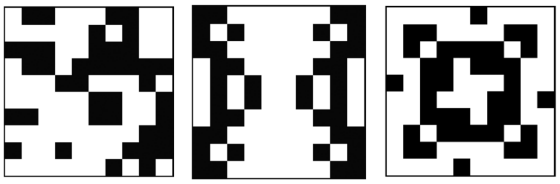
This domain covers a field of study within psychology and neuroscience where perception is studied using technologies such as eye tracking**Younes2016** and leveraging environmental controls in experiments where subjects rank images. There is some overlap within the field of IAQA and its dataset, such as the Waterloo IAA**Liu2017a**, where IAQA has used techniques normally applied within psychological experiments[2].

[2] A frequent critique of many IAQA datasets lack of control of ground truth data generation.

(a) Curve 2D Shape Line **(b)** Non Curve 2D Shape Line **(c)** Curve Couch Design **(d)** Non-Curved Couch Design

Figure 1.3: Examples of Visual preference (a)(c) vs less pleasing (b)(d) **Bar2006**

This is a relevant field, as it has provided controls such as screen calibration, image display, subject demographic, and randomization of stimuli**Leder2019**, **Mullennix2013**, **Rolls2014**. Empirical aesthetics also provides insight into areas such as visual symmetry, which are known to have also been a subject of computer vision**Elawady2017**, **Gray2010**, and an understanding of how viewers might rank an image.



(a) Random Image **(b)** Symmetrical Image **(c)** Rotational Symmetry Image

Figure 1.4: Examples of images used in psychological experiments showing visual preference for symmetry and complexity **Bertamini2013**

These studies have shown experimentally both that formal rules such as the 'golden ratio' and 'averageness' in facial portraits (figure 1.5), symmetry (figure 1.4), complexity**Bertamini2013**, and curve (figure 1.3) are attributes that are associated with aesthetically pleasing design and images (figure 1.5).

Figure 1.5: Average Composite Face (left) , Golden Ratio (Middle Right) **Brielmann2018**

However, this is far from the whole 'picture' and it has been shown both in computer vision **Simond2015** and in imperial aesthetics publications alike that attributes such the golden ratio or rule of thirds in photography**Green2012** and symmetry**Leder2019** are not universal aesthetic attributes and depend on context - this underscores the complex nature inherent within the domain of aesthetics.

1.1.4 Art Aesthetics

Art aesthetics and criticism include consideration of brush strokes and artist techniques, alongside the interpretation of images within cultural and critical studies that intersect

with social sciences, literary criticisms and philosophy. *Perspectives* within art aesthetics might consider or include *Marxist, psychoanalytical, feminist (figure 1.6) and queer (figure 1.7a)*.

The tradition in the 20th century has been to apply intellectual theory that has also been applied to other art forms, such as film and literature, and is not limited to considerations of *good* or *bad* opinion but rather critical and conceptual analysis using vernacular such as 'feminist aesthetics' or 'feminist critique'**Hein1990** of an artwork, item of popular culture or an image.

One fascinating feature of 20th century art is the inclusion of critical theory into the artwork aesthetics itself and 'aesthetics' that are ironic or a pastiche of popular culture's aesthetics (figure 1.6 shows two renowned and striking examples of this).

(a) (sic.) Untitled (Your Body is a Battleground)[3] **Kruger1898**

(b) Do Women Have To Be Naked To Get Into the Met. Museum? **TheGorillaGirls1989**

Figure 1.6: Examples of Late 20th century Feminist Artworks

Art aesthetics this are important and relevant in considering the possible applications of IAQA and given how different models that are trained apparently handle ambiguous examples and the potential to provide a 'critical lens' on areas such as cultural, racial or gender bias that might be replicated by machines. One might ask here how computer vision might learn feminist aesthetics?

Much of the focus of IAQA has been on what might fall within extraction of features: formalist art aesthetics, such as lighting, perspective or particular aesthetic techniques such as chiaroscuro (using lighter or darker tonal values to make something appear three dimensional)**Hobbs1990** shown in 1.7.

Another example of formalist aesthetics involves using attributes such as colour harmony - or using colours that are equidistant on a colour wheel (figure 1.8) - which positions discrete colours according to position, following prismatic splitting (figure 1.8a) (or electromagnetic spectrum) first theorized by Newton *red, orange, yellow, green, blue, indigo, violet*.

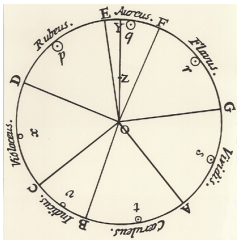
Many formalist attributes are used in hand crafted IAQA, and a consideration to wider aesthetics is included as there may be novel applications of computer vision to art criticism. One such high-level example of formal aesthetics is shown in figure 1.9; a focused example of features in IAQA literature is shown in table 1.1.

(a) Self Portrait by Artist Robert Mapplethorpe **Mappithorpe1980**

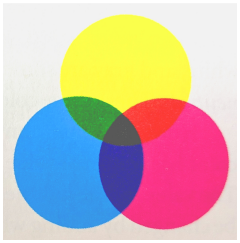
(b) El Fabula(an Alegory) **Domenikos1580**

(c) Salome Receives the Head of John the Baptist **Caravaggio1609**

Figure 1.7: Chiaroscuro examples



(a) Newton's Color Wheel Based on EMS (Prismatic Splitting)



(b) Primary and Secondary Hues in Pigments **Hobbs1990**

(c) Aesthetics Color Wheel **Hobbs1990**

Figure 1.8: Examples of Colour

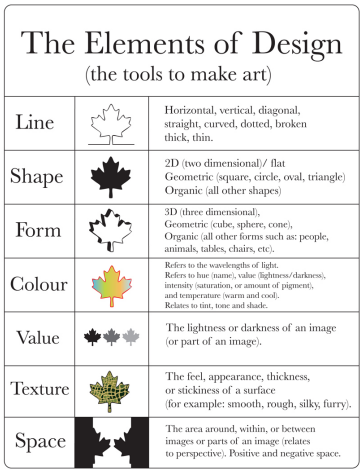


Figure 1.9: Elements of Design **Butler2012**

1.1.5 Photo Aesthetics

Photographic aesthetics within the field of IAQA have focused on so-called 'formal photographic rules', which overlap with more widely applied aesthetic rules, but have distinct properties which are a result of physical properties: lens (bokeh, Depth of Field(DOF), shutter speed), sensor, and light. Within traditional handcrafted approaches, this involved high-level attributes such as *simplicity, colourfulness, color combination(harmony), sharpness, image pattern, and object composition***Liu2017a**, **Mavridaki2015**, **Datta2006**, **Tang2013a**, **Simond2015**, **Lo2013**. These are informed by texts on photography that themselves frequently borrow from formalist art aesthetics, which have, in turn, been sources from which high-level features are defined.

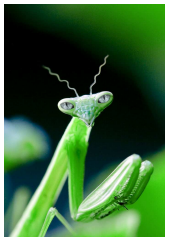
Others have sought to define more abstract features, such as spatial richness**Lo2013**

(a) Complimentary Colors(Color Contrast) Yellow/Blue **Ke2006**

(b) Colour Harmony **Lo2013**

(c) Shutter Speed **Lo2013**

(d) Rule of thirds **Mai2011a**



(e) Shallow Depth of Field (Blurring Background) **Ke2006**

(f) Good Lighting **Lo2013**

Figure 1.10: Examples of Photographic Formal Aesthetics

and similarity measure**Datta2006**; aesthetic attributes that are concepts of a hypothetical viewer mapped onto ranking measures. Table 1.1 shows attributes frequently used in IAQA literature, many of which are defined within photographic texts, photo guides or researchers' own experience. It is also clear from, for instance, attributes like *good lighting* are not necessarily easy to define though they may appear obvious, and that these include detailed aspects such as chiaroscuro, dynamic range, and contrast.

Further, these attributes are not strictly cumulative and/or necessarily combinatorial - attributes that may constitute a *good* photograph for one subject would not necessarily be effective should the subject be elsewhere. There are numerous examples of this in the datasets section of the appendix **??**.

However, it is also clear that having an intuition as a photographer is part of the aesthetic toolkit, and a re-combinatorial approach - using the right effects alongside the adjustment of camera settings, lense choice - is a part of what makes 'good photography'.

Figure 1.10d bottom clearly has *rhythm, colour harmony, composition* created by the object of the photograph being at the intersection of the golden ratio, the colour of the leaves, the repetition of trees and it also makes use of perspective - an image figure 1.10f shallow DOF(wide lens iris aperture/low f number) would be not be a useful attribute and would render the bridge out of focus. Attributes such as shallow DOF that in one context (portraiture or macro photography) are part of highlighting a salient object that in another (landscape photography) would degrade the image aesthetics quality[4]. Many of the low quality images in the qualitative examples of IAQA datasets (section **??** of the appendix) show images that might otherwise be high quality but the image has of

[4] f 64 is a renowned landscape movement which makes use of extremely small apertures-rendering almost everything in focus to create a painterly quality of image, examples of this can be seen on the Metropolitan Museum of Art's web-page: https://www.metmuseum.org/toah/hd/64/hd_f64.htm

Example Aesthetic Attributes		
Feature(aesthetic Attribute)	**Figure**	**Description**
Rule of Thirds/ Object Composition	**1.10d**	Salient or important elements placed on lines of image divided into 3×3 grid. Ratio of 1.61803 **Yeh2012**
Saturation/ Hue Saturation	**1.10a**	saturation indicate chromatic purity (emphasis of single color value) **Datta2006**
Simplicity	**1.10a**	Keeping single subject or saline object well composes and not having too many subjects in frame.**Tang2013a**
Color Combination	**1.10b**	Color Distribution of image finding a few dominant colours**Lo2013**, **Tang2013a**
Lighting	**1.10f**	lighting contrast between foreground and background, use of effects such contrast**Kong2016**, **Lo2013**
Sharpness	**1.10e**	Subject in focus, use of depth of field and lighting to isolate salient object from background**Mavridaki2015a**
Image Pattern	**1.10d bottom**	Use of symmetry and texture to crate harmony and rhythm**Mavridaki2015a**, **Kanwal2021**

Table 1.1: Aesthetic Attributes from IAQA Literature

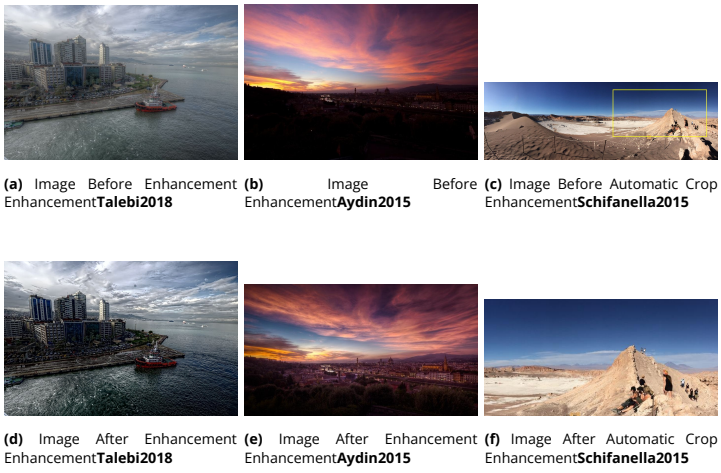
shallow DOF with a salient object that is out of focus.

One further aspect of photo aesthetics that is important to consider is the properties of the camera, this might include resolution of sensor, bit depth of color, alongside the sharpness of lens which effects both features such as chromatic aberrations (red, blue halo at object edge) and resolving power of lenses, many of these properties are interrelated. For instance small f number (aperture) increase sharpness even for low quality lenses but also increase DOF. Other aspects include tilt and shift (figure **??**) where a specialist lens or large format camera might be used in architecture photography to correct perspective.

1.2 Applications

1.2.1 Commercial Applications and Practical Applications

Applications include image recommendation, photo album optimisation**Liu2017** and advertising, alongside image enhancement**Talebi2018** and on-device real time IAQA. Others have utilised image captioned ground truth to train models that have then been able to offer advice to photographers, providing feedback on aspects of technique and camera operation such as,*'a greater depth of field of say f8 would have produced an over all better shot'***Jin2019**, these latter approaches have also made use of attributes' radar maps**Jin2019**. Another related approach is 'best hand shot'**Schwarz2018a** which utilises the multi-capture feature that many smart phones are now equipped with.

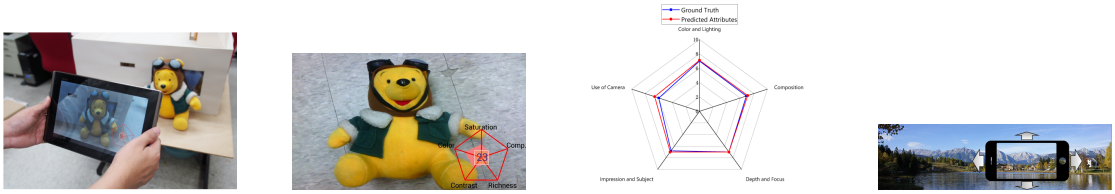


(a) Image Before Enhancement Enhancement**Talebi2018** **(b)** Image Before Enhancement**Aydin2015** **(c)** Image Before Automatic Crop Enhancement**Schifanella2015**

(d) Image After Enhancement Enhancement**Talebi2018** **(e)** Image After Enhancement Enhancement**Aydin2015** **(f)** Image After Automatic Crop Enhancement**Schifanella2015**

Figure 1.11: Image Enhancement Examples From IAQA Literature

Other applications include the selection of product photographs that are most likely to be of higher quality and hence more attractive to consumers, or of the most appealing architectural photographs of hotels. One real-world example of this is Idealo**idealo2021**, who provide a service and commercial applications of IAQA**Talebi2018**, **Lennan2018**.



(a) On Device Rating**Lo2013** **(b)** Aesthetic Attributes**Lo2013** **(c)** Predicted Attributes Scores (Radar Maps)**Jin2019** **(d)** Best Handy Shot**Schwarz2018a**

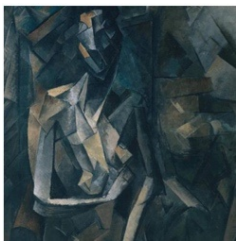
Figure 1.12: Application Examples (On Device Rating, and Best Hand Shot)

More widely, there are numerous fascinating examples of computer vision in aesthetics - such as pose analysis of ancient art, and art restoration and verification where computer vision is being applied to verify a painting's authenticity[5]**Snow2017** and in generating

[5] https://art-recognition.com/

novel images[6] and include computer vision applications of style recognition**Cheng2021** shown in figures (1.13a, 1.13b, 1.13c).

(a) Original Image **Cheng2021**



(b) Style **Cheng2021**

(c) Style Transfer Image **Cheng2021**

Figure 1.13: Examples of Style Transfer

[6] https://www.artbreeder.com/

1.3 Aesthetics: A Computer Vision Problem

Computer vision has focused applications stenography **Das2021**, texture analysis,semantic segmentation**Shelhamer2017**, 3D scene reconstruction**Murez2020**, super-resolution**Liang2021** and edge detection**Poma2021** as well as more widely in areas such as object recognition, image captioning, medical imaging. Prior to deep learning approaches, many techniques involved extracting features using Discreet Wavelet Transoms (DWT)**Makbol2013**, **Goel2019**, **Kanwal2021**, **Goel2014** and handcrafted or combinations of low and high level feature extraction, where support vector machines (SVMs) are used to enhance for multi-class or object recognition**Makili2011**, **Goos2016**.

While many of these methods continue to be researched, more recent approaches use Deep Learning - such as attention based salience**Zhu2020**, object detection**Mutch2006**, **Lei2019**, **Peng2018**, classification **Jia2020**, and multi-class or active learning**Wu2020a**, **Joshi2010**, **Li2004**, **Gu2015**.

Aesthetics as computer vision problem is further addressed in the Methodology Chapter **??**, and approaches outlined in the Literature Review **??** might be considered to be at the intersection of empirical study in the aesthetics of digital photographs (DP), Image Processing (IP), and machine learning (ML).

DP are quantized $\mathbb{R}2$ discrete time images of continuous signal data $\{a,b\} \implies \{x \in \mathbb{Z} : a \leq x \leq b\}$ where $a = 0.4 \times 10^{-4}$ and $b = 0.7 \times 10^{-6}$ frequencies of the electromagnetic spectrum in continuous time in $\mathbb{R}3$ space. IP is a set of techniques used in processing digital images; Machine Learning is a wider umbrella term encompassing approaches such as Deep Learning and Reinforcement Learning (RL). Approaches to deep learning include Convolutional Neural Networks (CNNs)**LeCun1989**, Generative Adversarial Network (GANs)**Goodfellow2020**, Long-Short Term Memory (LSTM)**Hochreiter1997a** and most recently Vision Transformers (ViTs)**Yuan2021**, **Dosovitskiy2020**. Each field $\in \{DP, IP, ML\}$ (an illustration of this is shown by the Venn diagram in figure 1.14) in turn might have further subsets, such as deep learned features as a supervised learning task on a particular subset of DP images, where one type of IP is used for as part of data augmentation. **morris2004computer**.

Many of the traditional approaches to aesthetics have relied on the extraction of hand-crafted features, and are what art criticism would define as formalist aesthetics; texts on formal photography such as **Kodak1995take**, are frequently cited in computer vision research**Mitarai2013**, **Murray2012**, **Talebi2018**.

One clear and obvious drawback of hand-crafted feature extraction approaches, in light of findings across subdomains of aesthetics (empirical aesthetics, philosophy and art), which has been noted in many of the computer science publications**Zhang2021d**, **Ke2006**, **Kolesnikov2020**, **Lu2015b**, **Birkhoff1933a**, **Brielmann2018**, **DaSilva2017**, **Skov2019** is the not unambiguous in definition(s) of attributes in compounded further by the potential for ambiguity within GT data. Further, although hand-crafted features have produced models that have been able to generalise, they have often been weak learners. Irrespec-

tive of recent gains made in computer vision using deep learning, which has come to dominate**Zhang2021d**, it might be argued that these later approaches are better suited to this type of domain - as features are learned, and therefore do not rely on human defined categories but rather on how features that best predict ground truth (GT) labels.

1.3.1 Image Aesthetic Quality Assessment IAQA

One conceptual challenge in defining a domain within aesthetics or vision is the question of whether aesthetic appreciation is cognitive, a property of stimulus itself, or something that can be objectively assessed as a quantitative and computational task.

This ambiguity has been highlighted even in early canonical attempts at analysis of aesthetics; Kant in 1892, for instance, introduces in his critique of judgement the notion simultaneously that 'beauty is in the eye of the beholder'**Zhang2019** while judging something as good - is to 'to make [...] a subjective assessment [of what] he has reason for expecting a similar delight from everyone'**Kant1892kant**. The tension between these two statements being an attribute of 'aesthetic' experience**Spratt2015**.

Notwithstanding these inherent ambiguities which are nontrivial, IAQA *has* been extensively researched and formulated as a computer vision problem. Addressing the question of where IQAQ 'sits', however, is important in dealing conceptual ambiguity. The importance of clarifying IAQA is fourfold:

Enables formulating a problem by understanding of what tools are appropriately leveraged from machine learning and pattern recognition;

Disambiguates IAQA within nomenclature and sub/super domains where there there is potential for cross over and conceptual confusion;

Identification of opportunities for real world application, learning for other domains and clearly;

Understanding overall purpose and value.

IAQA definition

Here the term IAQA is used to disambiguate the field of enquiry, as a term most frequently used in literary criticism and to disambiguate from from fields such as Image Restoration or IQA.

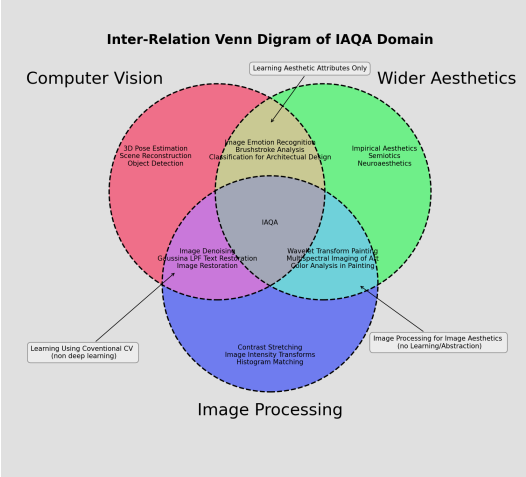


Figure 1.14: Venn diagram Data sets Reviewed in literature reviews

Both Image Processing and Computer Vision are included as separate fields, as there are applications for each within the wider domain of aesthetics - for example, Wavelet Entropy for painting crack identification does not necessarily constitute learning. Here, therefore, it is considered a qualifying feature if IAQA is an application of both Image Processing and Computer vision in the domain of aesthetics**Constatantin2020**.

Attention Mechanisms

Within the domain of IAQA, there remains a significant gap between hand-crafted or deep learning models in predicting how a population of human assessors would rate an image. Further, it is notable that the largest recent gains in predictive accuracy and ability to generalise have come from models that combine deep learning with attention mechanisms.

Attention maps, for example, have been framed as optimisation models to boost a model's ability to learn, to use semantic information, identify areas of focus, or using models trained on salient object detection to inform attention patch centroids.

1.3.2 Challenges for IAQA:

Some of the technical challenges that have been demonstrated to hinder a model's ability to learn are also useful areas for domains more widely. For example image composition is a feature which is challenging to learn where input images and corresponding feature maps of a convolution neural network are necessarily square. These are inherent and no trivial examples of how a CNN might handle real world data, which is potentially messy, ambiguous, and degraded or noisy signal data.

GT is often created by a community of human participants, some of whom may be hobby photographers and others professional. It is unclear whether these rules have been rigorously followed in all cases. Further, in classification tasks or object recognition salience detection, the levels of semantic granularity applied is not so well-defined for IAQA.

One area of complexity is the clear interplay between context and what is an aesthetic

attribute of an image - for instance, a medical image may be considered aesthetically beautiful, or indeed an image from the Oxford flower dataset, and if entered into a competition may score highly.

This may be, for instance, due to the subject matter having a high degree of visual interest that subjectively piques the interest of a viewer with attributes such as visual symmetry. It is clear that both image symmetry and symmetry of a salient object *itself* provide both potential attributes that can be considered aesthetically high value.

1.3.3 Challenges/Problem Definition

Advances in AI in the last decade have seen super-human performance in areas such as deep reinforcement learning and domains with high dimensional search spaces, such as protein folding**Senior2020** and the game of Go**Silver2016**) with its 10^{170} possible board combinations. Digital images provide one such example of high dimensional spaces, where an 8-bit 16×16 Red Green Blue (RGB) channel image has $x \in \mathbb{Z} \times 2^{8^{(16\times16)}}$ possible pixel combinations.

Within the wider field of computer vision, the use of deep learning techniques such as convolutional neural networks and transformers in Natural Language Processing (NLP) (and latterly, transformers for computer vision tasks) has resulted in performance accuracy that has increased year on year**Sermanet2014**, **Simonyan2015a**, **Szegedy2015**, **He2016**.

However, the most recent research papers published on IAQA in 2021 have not achieved accurate scores above 85% on benchmark datasets. Many of the approaches that have improved CNNs' ability to converge faster and generalise, better such as deeper networks **Simonyan2015a**, adding residual layers**Krizhevsky2017a**, and improving CNN optimisation algorithms**Kingma2015** have seen incremental minor improvements to IAQA.

Significant improvements have generally required attention mechanisms that introduce high-level spatial inductive biases during the training of combined widening or parallel multi-column networks. Many of these represent a replacement of hand-crafted feature extraction, with crafted policies/meta heuristic inductive biases on multi column architectures.

With recent metrics for classification CIFAR10**Dosovitskiy2020** >99% accuracy and CIFAR 100**Foret2020** 96.08%, image-net 90.88%**Dai2021**. Notably, **Dai2021**, **Dosovitskiy2020** are both convolution transformer hybrid (ConViT) models. While **Foret2020** is not a ViT-hybrid, it is only marginally better in performance than its next nearest neighbour **Dosovitskiy2020**. These metrics are so accurate that it would be important to ask if an individual human subject would be able to perform so well. Many of these improvements remain within narrower applications, where CNN's have already achieved high levels of accuracy and have remained dominant**Krizhevsky2017a**, **LeCun1998**, **He2016a**.

ViTs, to our knowledge, have not been used for IAQA, have traditionally required huge

datasets**DAscoli2021**, **Touvron2020a**, **Khan2021**, and only generalise well on datasets (14M-300M) images**Dosovitskiy2020**.

Datasets of this size do not exist for IAQA. Recent approaches using transfer learning are producing data efficient transformers**DAscoli2021**, however it is unclear, when compared with conventional approaches on smaller tasks and on new domains, whether ViTs or ConViT's outperform CNNs.

Here, we will *compare* and *contrast* the best performing CNNs with ViTs and ConViTs, and evaluate performance within the IAQA domain with binary (high, low) quality images, and attempt to predict how a compute of online voters would score an image.

1.4 Aims and Objectives

The objective of the dissertation project outlined here is to review IAQA using the AVA benchmarking dataset as a binary classification problem. This will be outlined within the context of wider datasets and implementation challenges, and the training of state of the art models using domain adaptation vision transformers and will outline what makes IAQA important.

Key aims and objectives are:

1. **Aims:** The aims of a this dissertation project are:
    - To build an understanding of previous computer vision research in the domain of IAQA;
    - To compare pre-existing CNN architectures with Vision Transformers (ViTs) and ConViTs;
    - To further knowledge on what state of the art models can contributions in the domain of IAQA;
    - To train, adapt and test*existing models* using available code;
    - To identify areas for future development;
    - To generate new learning by training models that have not yet been used within IAQA such as Vision Transformers.
2. **Objectives:**
    - To evaluate different ViT, ConViT and CNN models;
    - To improve models by adjusting hyper perimeters and selection of appropriate data augmentation;
    - To address challenges inherent withing the data such as *class imbalance*;
    - To report metrics on architecture performance of base-line and state of the art approaches;
    - To produce trained models that can be used for inference on unseen data;
    - To provide provide analysis of models using both quantitative and qualitative results;

- To provide analysis and rational for the selection of models;
- To report results reproduced in by existing research withing the domain of IAQA.

1.5 Contributions

The contributions made here are chiefly of a novel approach to IAQA using ViTs and ConViTs (the first application of ViTs and ConViTs) to the field of IAQA. We demonstrate this contribution by:

1. Comparing side by side with CNNs;
2. Show both quantitative and qualitative results of inference;
3. Conducting analyse of how both transfer learning can be used with minimal additional training.

Within literature review we conduct a review of datasets on IAQA and conduct analysis of dataset sizes and type and produce a proposed data dictionary which would be us full for further IAQA research, we also aggregate dataset size and type which is currently missing from IAQA literature and hope that this will be useful for further research within IAQA.

We also web scrape and produce a data dictionary of all 320k images available on dpchallenge.com which would facilitate further research including multimodal research an make use of the rich comments data that is available on dp.challange.com.

We also conduct analyse of Ava Benchmark Dataset and show that their is a statistically significant relationship between the competition and mean observed score (MOS) where competition number increases monotonically with time.

1.6 Thesis Organisation

The thesis introduction (chapter 1) has provided a high-level overview of the IAQA domain and potential applications. It will also introduce the current context, challenges and problems alongside providing context on aesthetics.

The literature review (chapter **??**) will provide context on both datasets and provide insights into how different publications have address the challenges of CV in IAQA; further it will review available data an provide justification for using the AVA Benchmark dataset based based on available literature and also outline evaluation metrics used.

Research methodology (chapter **??**) will outline the approach taken to data handling, the image processing pipeline, and to the training of models. Results and discussion (chapter **??**) will provide both qualitative and qualitative outcomes of the various CNNs and ViTs that were trained.

Chapter **??** provides future recommendations and outlines major contributions and the significance of the findings.

1. **Key Aspects:** within 1 *Introduction*.
    - Introduce the application domain of IAQA *aesthetics*;
    - Provide an initial outline of the problem and state of the art metrics(SoTA) on image classification;
    - provide a *(*lens) through which to read the following chapters.
2. **Key Aspects:** within **??** *Literature Review*.
    - A comprehensive overview of IAQA literature and analysis on different approaches taken;
    - Evaluation of what was successful and what any drawbacks and pitfalls were within the approaches taken so far;
    - Provide insights into how findings here can contribute to learning beyond simply attempting to supersede state of the art accuracy on the AVA benchmarking dataset.
3. **Key Aspects:** within **??** *Research Methodology*.
    - Vision Transformers overview and adaptation and outline models used and under what settings;
    - Address challenges posed by the data itself such as class imbalance;
    - Improve model performance using novel data augmentation techniques;
4. **Key Aspects:** within **??** *Results and Discussion*.
    - Evaluate wider significance of research approaches including consideration of cognition and where IAQA might contribute to wider understanding within the field of computer vision;
    - Evaluation of what was successful and what any drawbacks and pitfalls where within the approaches to IAQA thus far taken;
    - Provide analysis and rational for selection of models used.
5. **Key Aspects:** within **??** *Conclusion* .
    - Explore potential future research opportunities;
    - Outline contribution how this dissertation has contributed to existing knowledge;
    - Identify potential novel applications, significance and importance of IAQA and computer vision research.

The problem statement of binary classification on the AVA benchmarking dataset is outlined in the introduction and referenced, built on, and evaluated throughout the dissertation.

2 Literature Review

This section covers the baseline (section **??**) and state of the art models (section **??**) used within IAQA literature makes a definition of the data used and reviews various Hand-Crafted (HC) as well as deep learning approaches to IAQA. Within deep learning we also include (Vision Transformers) ViTs and (Convolutional Vision Transformers) ConVits and examine different attention mechanism used in DNNs (section **??**). In (section **??**) we make a case for deep learning over HC feature extraction and the reason for using the Aesthetic Visual Analysis (AVA) benchmarking dataset. We do this through conducting systematic review of datasets and conducting meta analysis of literature through the *'lense'* of datasets used in IAQA literature and descriptive as well as parametric analysis of the AVA dataset meta data (section **??**). Finally we outline the evaluation metrics used for IAQA as a binary classification problem in section **??**.

2.1 Related Work (Image Aesthetic Quality Assessment)

Image Aesthetic Quality Assessment (IAQA)[1] is a highly nuanced and challenging area, which requires care and consideration at several layers of detail.

At a high level within the literature on IAQA, there has been a shift from HC to deep features, in addition to the consideration of either visual attention mechanisms or model architectures. This is reflective of the wider trend in computer vision. Further to what is outlined in the introduction, here we narrow the focus to intra-IAQA literature, rather than interrelated sub-fields or domains that focus on painting or calligraphy**Fernando2021**, **Sun2015** etc.

There are, to date, four reviews (experimental and literature) on IAQA**Yang2019**, **Kanwal2021**, **Deng2017**, **Spathis2016**. These provide an initial high-level view into IAQA, and cover approaches both HC and many - but not all - of the datasets used in IAQA research publications. The prevalence of citation and review of various IAQA datasets is shown in the Venn diagram (figure **??**).

2.1.1 Baseline

Initial approaches taken to HC features Trainee Classifier on Support Vector Machines (SVM) or Gaussian Mixture Models: Here, we consider approaches that have used deep learning, alongside performance within IAQA rather than classification more widely.

Deep learning models have generally been binary classification, where MOS is thresholded $< 5 = 0$ and $> 5 = 1$ to produce binary ground truth of $\in \{0, 1\} = \in \{bad, good\}$

[1] Within the literature on image quality, some publications - which are clearly within the domain of IAQA - refer to this simply as Quality Assessment (QA)**Chang2017**; others heavily blur these categories**Kanwal2021**. For disambiguation, here it is referred to as IAQA so as to delineate the field from other domains - such as image restoration - clearly.

image quality.

As has already been outlined, some form of attention mechanism, paralleling, or patching has been used to improve results on various vanilla CNNs that have been used for IAQA as an image classification problem:

Alexnet **Krizhevsky2017a** by **Koa2016b**;

VGG16 **Simonyan2015a** by **Ma2017**, **Deng2017**, **Mai2016**, **Hosu2019**, **Sheng2018**, **Ma2017**, **Kong2016**, **Liu2017**, **Koa2016b**;

Inception Nets **Szegedy2016** by **Liu2020a**, **Talebi2018**;

GoogleNet **Szegedy2015** by **Jin2019**, **Hii2017a**;

DenseNet **Huang2017** by **Liu2020**, **Liu2020a**;

MobileNet **Howard2017** by **Talebi2018**;

ResNets **He2016a** $\in \{18, 34, 50, 101, 152\}$ by **Sheng2018**, **Kong2016**, **Koa2016b**, **She_2021_CVPR**, **Chen2020b**, **Liu2020a**.

These report an overall accuracy of VGG with AVA zero padded to square of 72.9%**Ma2017**, which is significantly less than the performance of ResNet152, reproduced here at 74.1%. Further, when comparing hybrid vision transformers *ConViTs*, many publications take ResNet architecture as baseline**Wu2021**, **El-Nouby2021a**, **Khan2021**, with more recent publications using ResNet152**Jin2019**. Here, training ResNet $\in \{18, 50, 152\}$ to produce baseline metrics.

Figure **??** shows one residual block, of which 18 make up the smaller ResNet figure **??**. These networks perform well, and are arguably more competitively efficient with fewer floating point operations (FLOPS). Each block has two weight layers, allowing for an x or identity layer to be concatenated back (a technique for countering vanishing gradients/keeping residual values). All of these networks are trained as baselines without attention mechanisms, but do not require consideration as far as feature hierarchy as this is a built in feature of CNNs**He2016** . Generally within IAQA, where an architecture is used to provide baseline metric, it is also used within a model as a backbone - although, this is not always the case. Authors have taken various approaches to copping, such as canter copping or warping to dimension $3 \times 224 \times 224$.

$$y = f(x\{W_i\}) + x \tag{2.1}$$

Where the output of one residual block x is input W_i are weights of i_th layer in the residual block.

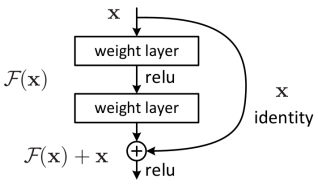


Figure 2.1: ResNet Block **He2016**

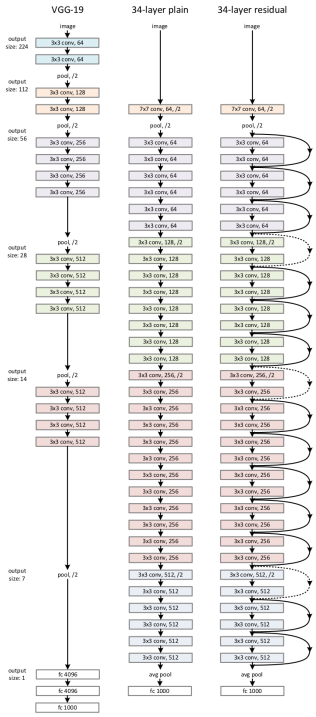


Figure 2.2: VGG-19 left , 34 Layer plain middle, ResNet 18 (19 blocks) right **He2016a**

2.1.2 Digital Images

All of the IAQA techniques reviewed are, by necessity, digital images produced using the sensor arrays of commercially available digital cameras (whose sensors are sensitive to light within an extremely narrow section of the electromagnetic spectrum of 0.4×10^{-4} (violet) and 0.7×10^{-6} (red)) and digital images discrete quantized representations 0-255 or $0 < (2^8) - 1$ of this spectrum.

Colour digital images are captured as three channels in an *RGB* image; traditional technique applies functions across an image array $f(x, y)$ where x and y are spatial coordinates of a pixel array.

2.1.3 State of the Art (SoTA)

Three main and closely interrelated features exist for deep learning based models, *multi column*, *patching of image*, and *attention*. For clarity and focus, here we evaluate models that are compared on the AVA benchmark dataset, as it has become the convention within IAQA to benchmark on AVA dataset.

Multi Column

Almost all models have used some form of multi column image classification network as a backbone; the performance accuracy of each is shown in **??**.

Even where the multi patch is not apparent in name, this is embedded in the approach in some way. Early models, such as rapid **Lu2014a**, almost replicate the extraction features and constitute HC thinking, but as a *policy* for a paralleled CNN, where the focus has been to train to classify based only on 'textures' (and then implement a voting system**Lu2014a**, **Wang2016c**), the most basic of these simply bolt on an SVM classifier to label parts of the dataset $\in \{scene, object, texture\}$ and then pass these to a separate forward feed network. **Koa2016b**, **Sheng2018** have separated learning tasks where semantic information is learned and use this alongside aesthetic to enhance predictive ability by concatenating weights back using a Bayesian probability graph at the end of the model.

Others have defined columns that train on spatially local patches that are selected via an attention mechanism, frame patches as an optimisation system, or they create a multi-patch aggregation system **Lu2015b**, **Sheng2018**. The most complex of these are similar to policies**Sutton2018** within reinforcement learning (RL) and **Sheng2018** learning optimisation algorithm. For instance, this even appears to obey a Markovian property - further underscoring the parallel. However, authors do not themselves make this observation or frame learning in this way - perhaps it would have enriched their work if they had.

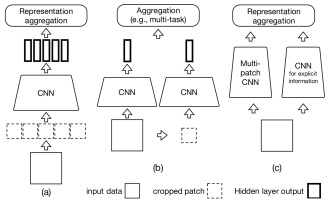


Figure 2.3: Aggregation based architectures for IAQA **Sheng2018**

Attention mechanisms

All of the state of the art models considered here are produced using attention mechanism or multi layer architectures that have been used to successfully improve classification accuracy, and produce models that perform better than vanilla networks on the AVA test set.

Some of these approaches, such as A-Lamp **Ma2017**, use adaptive patch selection based on the most informative patches, using graph based salience detection, to select multiple patches shown in figure **??**.

Various iterations of this have been the rule within recent IAQA literature, and these have been to solve a range of technical challenges, which fall into the categories of *spatial com positional, spatial intentional and semantic*.

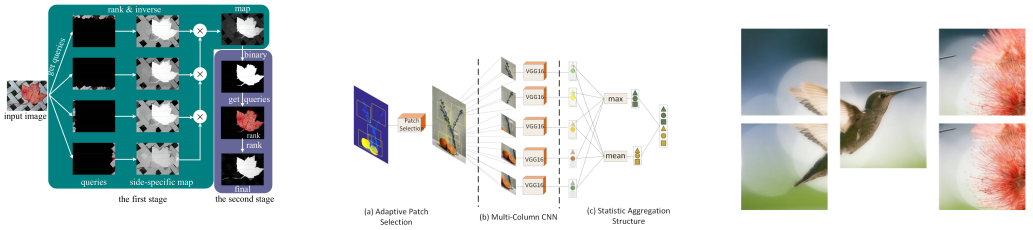
Spatially based attention is developed as a learning policy by **Sheng2018**, who employs an optimisation strategy to maximise training accuracy by selecting new patches and avoiding bounding box coordinates of the previous patch for n attempts. Other attempts have sought to create a policy where a model can focus on aesthetically similar images **Schwarz2018a**.

Some have cropped salient patches a without any other transforms and built an attribute relation graph **Ma2017**. The most recent and sophisticated of these create and utilise aesthetic related attributes and spatially related regions**She_2021_CVPR** or use multi-modal approaches**Zhang2021d** that leverage comments and information to combine self attention with attention based on LSTM trained on said comments.

The latter two approaches mark the the most significant improvements in overall accuracy in the AVA bench marking, and represent a leap of significant margin. One drawback of these approaches is that they do not necessarily lend themselves to easy application, and are difficult to validate. Further, each approach somewhat reinvents the wheel in terms of attention mechanisms; none of these approaches have used a vision transformer or hybrid *Convolutional Vision Transformer (CvT)* network.

Image Patching

The mean image size of the AVA dataset is 629×497 pixels with a maximum of 800^2 pixels. The convention in almost all of the approaches sub-sample, down-sample, or crop images as part of data augmentation. This is often the first process in image aug-



(a) Patching Example From Hand-Crafted Features **Yang2013**

(b) A-Lamp Multi Patch Network **Ma2017**

(c) Patches Taken From AVA Dataset Based on Saliency Graph **Ma2017**

Figure 2.4: Patching Examples from Hand-Crafted **Yan2013** and Deep Features **Ma2017**

mentation; image patches are in most cases $3 \times 224 \times 224$, however, in some cases images have been warped to square before patching **She_2021_CVPR** and others have been zero-padded to preserve composition.

State of Art Models and Metrics on AVA Dataset					
Year ↓	**Model**	**Reported Acc.** ↓	**Reproduced Acc.**	**Approach**	**Backbone**
2014	RAPID DCNN**Lu2014a**	75.4	–	Double column CNN (texture)	AlexNet
2015	DMA-Net **Lu2015b**	75.4	–	Multi-column	Alexnet
2016	A&C CNN**Kao2016**	74.5	–	Scene Object Texture (multi attr./col.)	Rapid CNN**Lu2014a**
2016	MNA-DCN**Mai2016**	77.1	–	Adaptive Spatial Pooling	VGG-16
2016	MNA-CNN-Scene**Mai2016**	77.4	–	Multi-Column (spatial)	VGG16
2016	BDN**Wang2016c**	78.9	–	Multi column (directed augmentation)	Custom Auto Encoder
2016	MTCNN **Koa2016b**	79.1	–	Baysian Multi Task	AlexNet, VVG Net, ResNet
2017	New-MP-Net**Ma2017**	81.8	–	Multi-Patch	VGG-16
2017	MultiGap**Hii2017a**	82.27	–	Multi pooled with text augmentation	GoogLeNet
2017	A-Lamp**Ma2017**	82.5	–	Adaptive-Multi patch (spatial)	VGG16
2018	NIMA **Talebi2018**	81.51	77.36	Reproduces MOS Distribution	Inception-v2, MobileNet
2018	MPADA**Sheng2018**	83.03	79.1	Attention Based Multy-Patch	Resnet18
2019	MLSP**Hosu2019**	81.72	81.7	Double column CNN	VGG-16
2020	AFDC**Chen2020b**	83.2	–	Muli-column adaptive mini-patch	ResNet50, VGG16
2020	FCN-A-G**Liu2020**	83.6	–	Region Graph Convolution (fully conv.)	Resnet101 FCN, VGG-16, DenseNet-121
2020	PA IAA**Li2020a**	83.7	–	Siamese Multi-column	Densnet121, Inception-V3
2020	MSCAN **Zhang2021d**	86.66	–	Multi Modal Self colab. attnt.	VGG16
2021	HLA-GCN**She_2021_CVPR**	84.6	–	Layout-Aware Graph CNN	Resnet-50, Resnet-101

Table 2.1: State of The Art Metrics (overall accuracy)

2.1.4 Hand-Crafted Features Genealogy

The approaches within IAQA fall into two main categories: hand-crafted (HC) and so-called deep features. HC features are extracted at both *lo* and *high* level and are extracted by traditional approaches to digital image processing**gonzalez2008digital**.

2.1.5 Feature Extraction

Early approaches to IAQA combined 'low-level' features and processes, such as wavelet transforms, to obtain image texture to 'extract' information, or applied coefficients recursively to obtain edge information, or applied some coefficient across an image to compute blurriness.

* Chronology of IAQA Techniques*							
Date	Technique	Feature Type	Extracted (imaging process)	Feature Levels ‡	Dataset	Metric	Score
2004	Ada-Boost, Real-AdaBoost (SVM, Bayesian) Classifier**Tong2004**	Texture, Shape, Color, Energy	Band Diff. color histogram, color moment, lab coherence, HSV Coherence, DFT Moment,DCT moment, Wavelet, hist $\in \{Sobele, Laplace, Canny\}$,	H/L	B†	Linear Corr.	84.7%
2006	SVM Binary Classification**Datta2006**	Texture, Familiarity, Size, Depth of field	Wavelet Based and 56 subset functions	H/L	P.Net	ACC	70.12%
2006	Naive Bayes class **Ke2006**	Glob Edge Distribution, Color Dist., Lo-level Contrast, Brightness indicator	Laplace Filter, KNN, 20bin Histogram, Fourier Transform	H/L	DPC	ACC	72%
2008	SVM, Gentle ADABoost, Bayes Classifiers**Lou2008**	clarity, contrast, lighting, simplicity, composition geometry, color harmony	Histogram, blurring kernel, $f(Hue \times Bri. \times Sat.)$	H/L	DPC	max AUC	93%
2009	SVM Classifier**Wong2009**	salient object, subject background	Saliency Map	H/L	P.Net	5CV-ACC	78.8
2011	SVM Classifier **Dhar2011**	Colour, DOF, Illumination of sky	Color spatial dist., multi scale contrast, Wavelet Energy, Haar Features, Spatial Pyramid Shape	H/L	DPC/Flkr.	ROC	–
2012	SVM Classifier**Lo2012a**	Layout, Composition, Texture, color	Hist, FFT, kNN(color)	H/L	CUHK	mean ACC	86%
2013	SVM Classifier**Tang2013a** (feature prediction)	Geometric composition, Complexity, DOF, Hue Composition, Blur, Brightness	Color harmony f, Orientation f, Salience map, Kernel Blurring	H/L	CUHKPQ	max AUC	$< 80\%$
2015	Bayesian Network Support Vector Regression (SVR)**Gao2015a**	17 High level	SSIM, SIFT, HOG	G	DPC/PNet	ACC	72.7
2015	SVM Classifier **Mavridaki2015**	Proposed Simplicity, Colorfulness, Sharpness, Pattern, Composition	Histogram Wavelet coefficients, Euclid distance, Color histogram	H/L	CUHK-PQ/AVA	ACC	77.1%
2016	SVM **Wu2016**	Aesthetic Class	mulitmodal,Structural Features, Local Vision features, Functional purpose	H	DS2	ACC	78.42

‡ High(H) Low(L) Features, † Bespoke

Table 2.2: Approaches to Feature Extraction

The trend in the use of HC features seems to be from multiple complex and individually defined functions, for example **Datta2006** and **Tong2004**, (early examples) both of whom extract 56 and 15 individual features respectively and later examples **Gao2015a** extract only three but apply a more complex learning algorithm.

Global/Local

Almost all approaches use both global features such as texture, and local features such as salient objects. There is some ambiguity in the literature about what constitutes 'local'. Examples of a low-level features are sub-banding in Wavelet, convolving across images, where there is therefor some spatial locality or granularity to the extracted feature.

Pipeline

Many of the features are highly complex functions that have relied on the knowledge of signal processing and making use of Fourier spectrum transforms**Ke2006**, as well as other are more simple histogram calculations (where pixel values are binned into hues)

to calculate color harmony.

Figure**??** shows a typical HC feature extract-on and SVM training pipeline.

Many of the HC feature extraction approaches reviewed here grouped various features into categories or types such as composition**Lou2008** or blurriness**Ke2006**. Extraction in the design is that of crafting, using engineering skill and domain knowledge to select the design features that correlate with image quality.

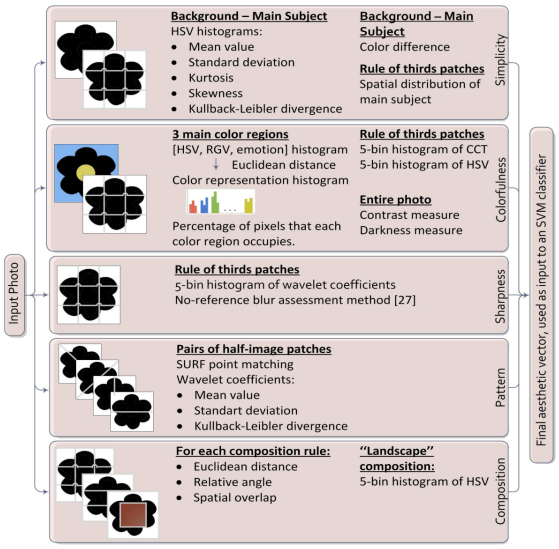


Figure 2.5: Hand-crafted feature extraction pipeline **Mavridaki2015**

Convolutionally *or* Conventionally Derived Features

Approaches, such as convolution operations, which were formerly a mode of implementing a process of applying a predefined coefficient, have shifted to become fundamental and dominant attributes of deep learning for computer vision.

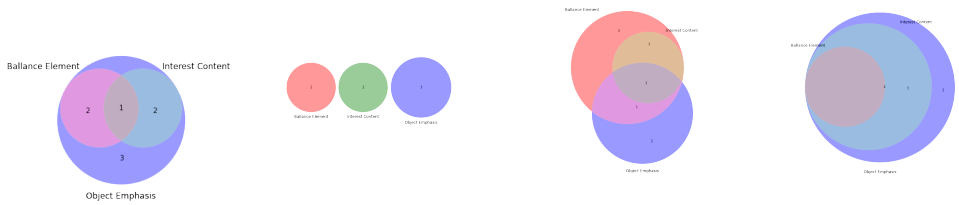
Deep learning inverts this, making the convolution operator central and neglecting such aspects as high or low-level features, domain expertise, or fundamental understanding of physical properties, which image processing requires. Within this, it is interesting to note that more recent HC feature driven approaches tend to favour simple features and complex image categories**Mavridaki2015**, **Gao2015a**, perhaps in response to deep learning's dominance?

Further, of the HC approaches comprehensively reviewed and shown in Table **??** only one compared results with the AVA dataset **Mavridaki2015** where side-by-side accuracy is 0.77% on a subset of only top and bottom 10%(score 1 and score 10) quality images(ignoring 80%) of data including most ambiguous images. The most accurate binary classification HC approaches were trained on Chinese Hong Kong University-Picture Quality (CUHK-PQ) dataset**Lo2012a**. This dataset is heavily annotated, and has had ex-

pert input in rating high and low quality, in addition to being much smaller. All approaches have performed well on this dataset of professional images.

2.1.6 A Case for Deep Learning

There is clearly a great deal of interlinking between categories, such as object emphasis, the 'rule of thirds', and shallow DOF, such that a point of focus or object emphasis might be achieved through rule of thirds or shallow depth of field which are categories defined, for example, in the AADB (Aesthetics Attributes Data Base) (section **??** figure **??**) with GT labels and HC features. Even 8 HC features can result in a great deal of complexity and ambiguity - not withstanding the high degree of human interpret ability of such concept as a 'balanced element'.



(a) $I \cup B \subseteq O$ **(b)** $I \notin B \notin O$ **(c)** $I \cup B \cup O$ **(d)** $I \subseteq B \subseteq O$

Figure 2.6: 4 Possible configurations of inter-dependant features within sample space of 8 possible features

This process as has a high-branching factory and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ possible subsets $\binom{n}{k}$ where n is the number of features n and k is number definable or possible features *feature space* that constitute a high quality image. In a search space of $x \in \mathbb{Z} \times 2^{8(n \times m)}$ this is clearly a high number.

This has the drawback of reliance on human pattern recognition and human defined features, to which machine learning is then applied. In removing the need to define features by hand *a priori*, deep learning represents a paradigm shift and has several advantages for the IAQA domain:

- Removal of human bias from feature definition, where definitions within aesthetics remain inherently ambiguous;
- Potential ability to generalise better;
- Finding patterns in image data that humans would not;
- Models can be used to predict image quality of new images easily.

However, while the techniques employed on smaller datasets, such as AADB, may generalise less well on unseen data, HC approaches are appropriate for generating learning from a smaller databases. And approaches that combine both, such as **Kong2016**, provide valuable insights for later analysis. These earlier forays into IAQA should not be disregarded.

2.2 Related Data sets

This section will provide an overview of the different datasets that have been used for IAQA. The purpose of this is to provide further detail, outline some of the inherent challenges, such as class balance within, and how to formulate a problem in such a way as to be able to effectively train and measure against benchmarks, and to illustrate the history and development of the field and draw out how new learning has been generated in previous work.

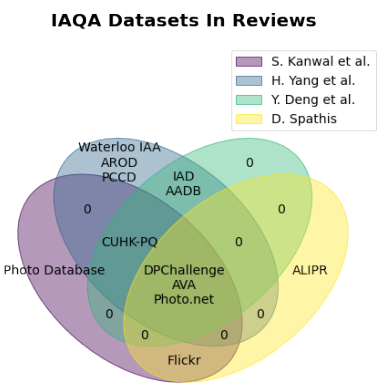


Figure 2.7: Venn diagram Datasets Reviewed in Literature Reviews

Within the literature that surveys the IAQA domain, there are 15 datatsets (shown in Table **??**) that are reviewed in some capacity, and have had publications associated with them. There are also a number of publications that have formed their own bespoke datasets, or which are subsets of existing dataset. In IAQA Reviews **Yang2019**, **Kanwal2021**, **Spathis2016**, **Deng2017**, there are 12 datasets covered. IAQA reviews consistently cover AVA and DPChallenge, however a significant number of datasets are only mentioned by a single source see **??**, this is reflective of wider IAQA research.

Although their are many IAQA datasets,it is clear from the variability both in overall size and type of annotation that, while benchmarking datasets in other areas of computer vision, such as image classification (Cifar $\in \{10, 1000\}$**Krizhevsky2009**, **Krizhevsky2009a**, ImageNet**Deng2009** $\in \{1K, 21K\}$), IAQA datasets are less well defined and readily available. Within this, several subset dataset have been formed to address challenges such a minority class or sparsity of a super set have been compiled.

Image Aesthetic Quality Assessment Data Sets					
Name	Size(n images)	Classes	Year	Additional	source
AADB**Kong2016**	10,000	2	2017	12 subsets	Flk.
AROD**Schwarz2018a**	304,000	2	2018		Flk.
ALIPR**Datta2008**	13,010	10	2018	sentiment	Flk.
AVA**Murray2012**	255,530	10,2	2012	discretized	DPC
CUHK-PQ**Tang2013a**	17,690	2	2013	7 sub-cats	DPC
DPChallenge**Datta2008**	16,509	10	2008		DPC
FCDB**Chen2017**	4135	∞	2017	cropping	Flk.
Flikr**Yin2012**, **Chang2017**	80000	2	2017	geo-tagged	Flk
Flickr AES **Ren2017**	40000	2	2012		Flk.
IAD**Lu2015a**	1,500,000	2	2014		Multi
IDEA**Jin2020**	9191	10	2020	ballanced	DPC
PCCD**Chang2017**	4135	7	2017		Bespoke
PD*Lo2013	1051	2	2013		CUHK-PK
Photo.net**Datta2006**	3581	7,2	2010		P.net
Waterloo IAA**Liu2017a**	1000	7	2017	discretized	P.net

Table 2.3: IAQA DATASETS

Figure **??** shows only three datasets of literature reviewers that are explicitly covered by

in all four IAQA review publications(AVA, DPChallenge, and Photo.Net). The AVA Datasets is a super set containing DPChallenge images and is clearly central to the IAQA domain.

IAQA ground truth scores are described through the literature as a distribution MOS (Mean Observed Score). Some datasets have fewer ground truth voters - 28 votes per image photo.net **Murray2012**, **Wu2011** and others have many raters from a single online community and have 210 per image (AVA); some are rated using a paid for service such as Amazon Mechanical Turk(AMT), or by specialist highly controlled informants where screen and demographic information are recorded in addition to image rating.

Dataset sizes range from 1100**Liu2017a** Waterloo IAA to 1.4 million**Lu2015a** images (IAD). The mean dataset size overview is 152k images. Figure **??** shows a plot of the dataset against the year of associated publication, and while there does appear to be small positive correlation, there is also an increasing degree of variability in dataset size, with both the smallest and largest datasets being first associated with publications only two years apart. Further, this divergence seems to increase with time. While available data in other

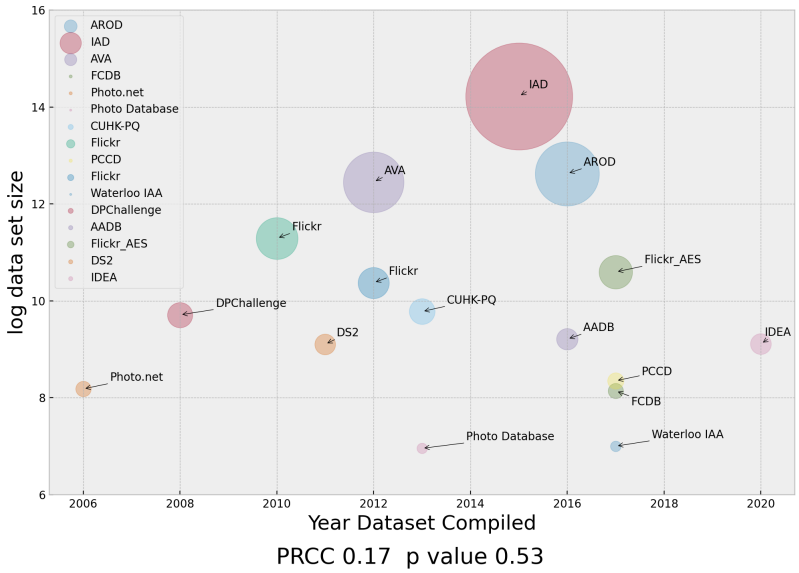


Figure 2.8: Dataset Size and Year of Publication Plot

areas of computer vision appears to have increased with time, this is not the case for IAQA, where even very recent publications propose novel datasets such WaterlooIAA**Liu2017**. There is no evidence to reject a null hypothesis that datasets do *not* increase over time.

The MOS distribution of IAQA datasets frequently appears Gaussian**Murray2012**, **Datta2008**, **Yang2019**, **Talebi2018**. AVA is the most normally distributed of these**Murray2012**, where

a Gaussian probability density function is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{2.2}$$

Where σ is standard deviation and μ is mean MOS. The AVA dataset has the most normally distributed dataset of any available IAQA dataset to date**Murray2012**, and has more than 300**AVA2012** associated publications. This is significant in particular for the AVA dataset, where the mean number of votes for each image $= 210$. This can be taken as a reliable ground truth value, where the probability of an image being scored on some scale is greatest at its mean and enables the exclusion of relatively few images as outliers from a normal distribution.

Further, it is significant that within this AVA Database of $\approx 255k$ images, only a very small number are excluded as outliers, where an outlier is $\pm 4\sigma$, which, while not in itself a test for a normal distribution, shows the normal test for an outlier is $\pm 2\sigma$.

This is consistent with the notion of so called *vox populi* or 'wisdom of crowds' and was initially researched and coined by F. **Galton1907Galton1907** in his significant demonstration that mean estimated values of crowds were frequently within $\pm 3.1\%$ of ground truth (GT)[2].

This significant finding provides a sound reasoning for aggregating crowd sourced scores as a measure of actual image accuracy. One aspect that is not taken into consideration with approximating GT, however, is online community bias, and one point of critique of the AVA dataset is that individual users are not identified and it is further not possible identify how individual users vote over time across different images, or conversely how an online community might vote for different images of the same user.

Classification tasks, which many IQAQ publications employ as routine throughout the literature, compute MOS of each image and then threshold this into usually good or bad image quality categories. This, however, results in class imbalance with the low category as a minority class. The practice of computing MOS (see eq. **??**) emerged in IQA where HC and traditional computer vision approaches such as**Wang2004** were employed as proxies for GT.

Figure **??**, left, overlays plots of gaussian distribution and the AVA dataset (light blue). The line's Probability Density Function (PDF) of a gaussian distribution e.q. **??** (red) and a computed PDF mapped to AVA MOS scores showing a slight positive skew. The main deviation from this is in values around the mean (note histogram bars above blue PDF).

One potential source of error aggregating human ratings is that it is not possible to rate 0, but it is possible to rate an image at 10. This may account for some psychological bias when rating images . Figure **?? right** appears corroborate this, with the positive scores

[2]Frances Galton would likely not have used the term ground truth

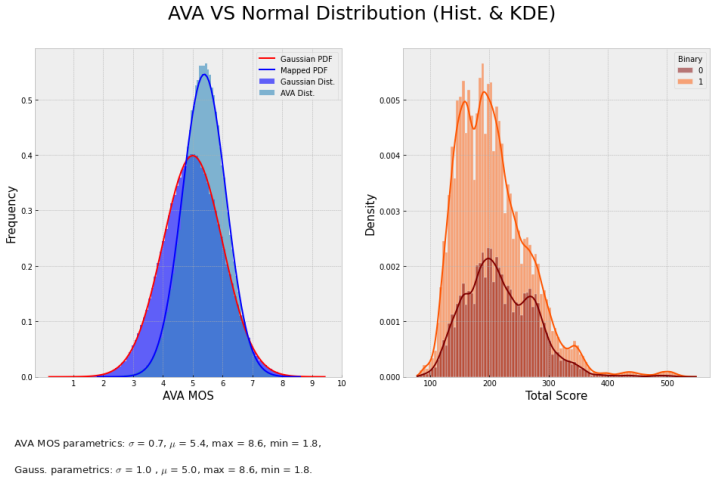


Figure 2.9: Left Compared Gaussian with AVA Distribution of MOS Scores with Gaussian PDF as Defined Above in **Right** and Kernel Density of Score: low (0) and high (1)

outnumbering the negative scores and also showing less normalcy of distribution. Figure **??** shows that as the number of votes increases, the average score appears to increase with it.

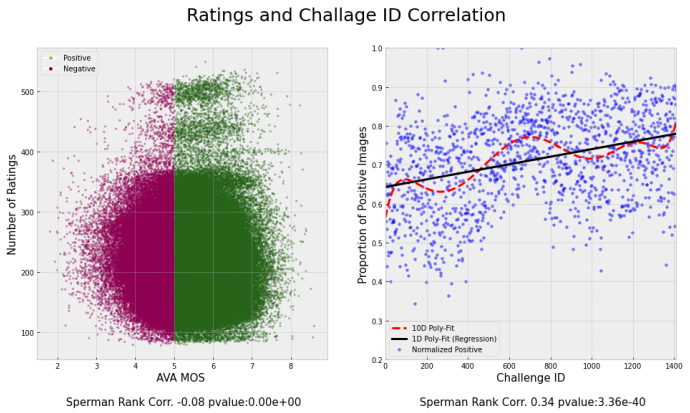


Figure 2.10: Number of Ratings MOS Correlation **left** Proportion Positive Correlation with Challenge ID ΔT **Right** (With 10 Dimensional Polynomial Fit(Red)

Figure **?? left** shows both class imbalance (greater high than low) and an apparent increasing proportion of positive score with time. While time is not recorded, with Challenge ID (which increases monotonically) each challenge is opened and closed for a discrete time period. Therefore, it can be shown that there is a positive correlation with change in ΔT **?? right** with Spearman's Rank Correlation Coefficient (SRCC) of 0.34 and p value of 3.36×10^{40} (this is almost certainly significant).

Accounting for this could take a variety of hypotheses, such as a developing online community within dp.challenge.com. It is certainly clear from reading the comments on dp.challenge.com that there is an apparent social network with many members being

persistent - this positive drift could be as a result of cognitive bias?

While some datasets, such as AVA, consist of 250k images (which may have been considered large at its inception), when compared with domains such as object recognition where models can be trained on datasets such as Image-Net (which contains 15 million images**He2015a**) the available datasets for IAQA remain relatively small. Within the wider field of IAQA, which includes moving images, quality assessment, and fascinating areas such as calligraphy.

Further, it is important to make a distinction between the field of image quality assessment (IQA) where benchmarking datasets such as LIVE IQA**Ghadiyaram2016** represent images that have been degraded by artificial noise and are used as benchmarks for the sub-field of image restoration within computer vision.

2.3 Evaluation Metrics

Almost all review approaches to IAQA are binary classification, with few treating as a ten class (reproducing probability distribution of images).

For classification accuracy, balanced accuracy, and F1 score are reported within the literature; 'accuracy' is the most frequently used metric in IAQA literature**Koa2016b**, **Schwarz2018a**, **Lu2014a**, **Ma2017**, **Chen2020b**, **Zhang2021d**. Much less frequent is balanced accuracy**Deng2017** and F1 score by **Ma2017**, **Mai2016a**. Many approaches use both regression and classification.

Accuracy

'Accuracy' is a measure widely used to evaluate model performance in classification tasks, on both evaluation and test sets.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

Balanced Accuracy

This metric is particularly important in evaluation during the training and testing of models where there is a class imbalance, which is a significant attribute of many IAQA datasets.

Balanced accuracy in the binary[3] case used within IAQA is given by computing the arithmetic mean of sensitivity and specificity False Positive Rate (FPR)(eq.**??**) (where sensitivity

[3]for non binary cases, class weights must be computed

or true positive rate (TPR) (eq.**??**):

$$Balanced = \frac{FPR + TPR}{N_c} \tag{2.4}$$

Where N_c is the number of classes,TPR is given by:

$$TPR = Sensitivity = \frac{TP}{TP + FN} \tag{2.5}$$

and FPR is given by:

$$FPR = 1 - Specificity = \frac{FP}{TN + FP} \tag{2.6}$$

Other approaches to addressing class imbalance during training and evaluation are discussed below in the Methodology section.

F1

F1 score produces the harmonic mean of precision (eq. **??**) and recall (eq. **??**) and is given by:

$$F1 = \frac{2(Precision)}{Precision + Recall} = \frac{2(TP)}{2(TP + FP + FN)} \tag{2.7}$$

where precision is given by:

$$Precision = \frac{TP}{TP + FP} \tag{2.8}$$

and recall is given by:

$$Recall = \frac{TP}{TP + FN} \tag{2.9}$$

The evaluation metrics outlined here are for binary classification, but a number of approaches train a regression problem first, or compute the mean softmax of ten classes (reproduced score distribution) and use Persons Linear Correlation(PLCC) eq. **??** Coefficient or Spearman's Rank Correlation Coefficient (SRCC) **??**.

This approach enables distribution analysis also, however it is not always clear in the literature whether a model has been trained purely as a regression model (reproducing MOS). Further, training a ten class problem is less challenging and does not give a good or bad quality estimation. This is particularly apparent with some of the more recent publications that achieve the highest accuracy training models with 10 node output of a fully connected layer.

In order to be able to produce an end to end network here, we train a binary classifier

rather than thresh-holding output computed MOS which approaches such as **Talebi2018**, **Zhang2021d** appear to take.

2.4 Literature Review Summary

We show that deep learning has surpassed HC feature extraction on large dataset where IAQA is a binary classification problem and outline why we conduct experiment on the AVA dataset. The role of image processing is now part of data augmentation, however many of the HC approaches provide valuable insights into IAQA features, in part as they require a human understanding of the domain. Deep learning inverts this, making convolution operator central rather than a method of feature extraction, secondly the process of making predictions beyond good or bad is more more straightforward with HC approaches where individual aesthetic attributes can be defined, extracted with a greater degree of explainability.

While many of the deep learning approaches outperform HC reliant feature extraction, deep learning approaches rely on hand crafting of complex learning policies on equally complex multi column network architectures.

We also show that ther are many IAQA dataset although many are overlapping and subset of larger datasets such as AVA. We also demonstrate that MOS scores approach a gaussian distribution and that there is a large majoriy class.

There is a great deal of literature available on IAQA and specifically using the AVA benchmarking dataset. There also exists some fragmentation within the field with many datasets where it is not always clear whether one is the subset of the other, further dataset metrics are not always reported consistently there are further example of this in the appendix **??**. This is further compounded by a slippage in nomenclature between data source and data set (DP.Challange). This ambiguity exists within publications **Talebi2018** use a 25k test set and **Hosu2019** use the 19k dataset which is from a test train validation split outlined by **Murray2012**. This is further compounded by the fact that the images are different in each subset making side by side comparison challenging.

3 Methodology

To our knowledge, the deep learning networks and architectures adaptations review in **??** only include traditional hand-crafted approaches and adaptions of various vanilla backbone architectures. Almost all approaches formulate problems as binary classifications and reproduce overall accuracy. There are two approaches to this: thresh-holded MOS prediction(regression) models output, and training pre-thresholded classes by GT MOS, both with a test train split applied by **Talebi2018**, **Murray2012**.

Here, we adopt the latter as a more challenging problem with a clear relationship to real world applications of IAQA and less room for ambiguity within the accuracy metrics (which the authors observe in some of the IAQA literature).

Some approaches appear to present the accuracy of a ten class (model) MOS, which will give higher accuracy on test data as the trained network will not have to tolerate fully the significant number of images with MOS very close to the high-low class threshold. Further, with a 2 node output vs 10 node there are quite simply 5 times less gradients to vanish.

Here, we train a various transformer models - both vision transformer ViTs and convolutional vision transformer CvTs. The reasons for this are:

1. Transformers have not yet been applied to IAQA and a side-by-side comparison generates new insights;
2. Many deep learning models have relied on hand-crafted attention mechanisms, a process which is a learned intrinsic feature of Vision Transformer(ViT);
3. Many more recent models have combined deep features with attention layers to obtain the best of both worlds;
4. There exist a number of pre-trained models on imagined allowing efficient transfer learning to and IAQA Domain.

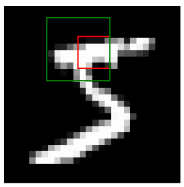
3.0.1 Side By Side Comparisons

Vision transformers require huge datasets and special treatment via custom training schedulers, which have warm-up and cool-down data phases.

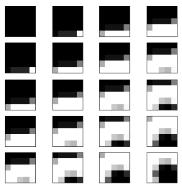
Comparing ViTs, CvTs and CNNs side-by-side on as close as possible training conditions enables an evaluation of whether the introduction of convolution is a means to soften the hard requirements of ViTs during training, providing inductive bias inherent in CNNs.

We reproduce an example to illustrate how CNNs convolve and maintain spatial proximity. Red shows the kernel and green bounding box shows the image area covered; here, kernel size and stride are parameters that effect feature maps learned.

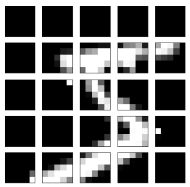
Contrast this with figure **??**: patches are exclusive and then attention is learned between *tokenized* discrete patches- clearly a very different paradigm.



(a) 5 from MNIST Handwriting Dataset **LeCun1998**



(b) Convolutions Filter Kernel of Size (5,5) Stride of (5,1)



(c) Patches Filter Kernel of Size (5,5) Stride (5,5)

Figure 3.1: Illustration of Spatial Inductive Bias

We employ domain adaption as an overall approach, where target domain is IAQA binary classification and source domain is multi-class classification convolution architectures of both convolution and ViT models. This brings with it the further strength of being able to use pre-trained models, which do not require compute intensive training from scratch. Transformers require large datasets**Zhang2021**, **Kolesnikov2020**, **DAscoli2021**, **Khan2021**, **xiao2021early**, **Wu2021** with then hundreds of millions of labelled data entries**Zhang2021** to converge and produce state of the art metrics, which is much larger than required for CNNs. This is at least in part because CNNs encode prior knowledge (adapting stride and kernal size) of image domain such as *translation equivariance* **Khan2021** across feature maps (figure **??** illustrates one example of this), CvTs do not have this as a possible prior embedding; these must be fully learned across discrete patches. Various approaches have been used to improve performance, such as introduction of gated position self-attention (GPSA)**Touvron2021a**, **Touvron2020a**, **DAscoli2021**. We will use domain, adaptation, and transfer learning as well as improvements to data requirements that have been introduced. This section will cover our approach on the following areas:

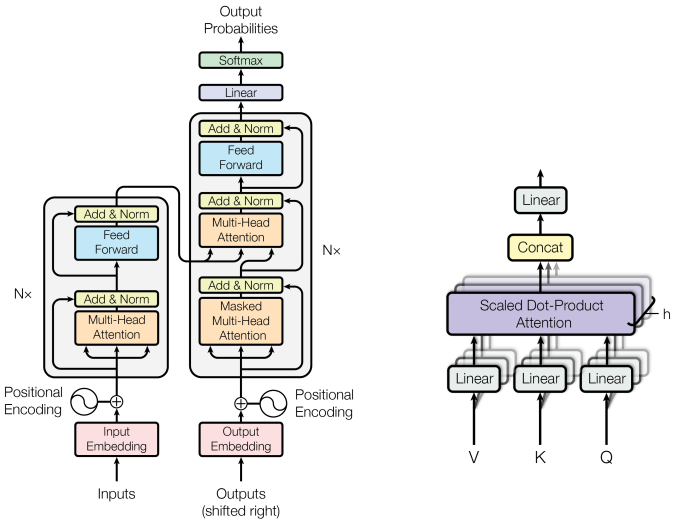
1. Transformer Architecture
2. Pipeline and Pre-Procession
3. Data Augmentation
4. Training Methodology
5. Testing Methodology

3.1 Transformers

Transformers have come to dominate over the recent years in many domains, have outperformed Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) such as Long and Short Term Memory (LSTM) architecture**Vaswani2017a** and have become a hot topic.

Many of these were designed to address challenges such as vanishing gradients, and in many areas, this began with applications such as machine translation within natural language processing (NLP)**Wolf2020a**. These were also areas where attention mechanisms have become an integral part**Vaswani2017a**.

The transformer consists of blocks (head) in which each learner in sequence learns align-



(a) Attention Model Architecture **Vaswani2017a** **(b)** Multi Head Attention **Vaswani2017a**

Figure 3.2: Initial Attention Model Diagram

ment in K, a value query system where input $Q$is passed to an embedding layer One-Hot tokenization. Input is a tensor of shape $\mathbb{R}^B \times \mathbb{R}^N \times \mathbb{R}^D$ where B is batch size, N is sequence length, and D is dimensional embedding.**Tay2020**. Each attention model architecture can be seen in figure **??**

Self Attention

Attention is is a mapping of query (Q), key (K), value (V) pairing - this is not unlike the python dictionary with a key store. V, K, Q are MxN square matrices. The matrix Q is the product of the computation of a set of queries with all keys**Vaswani2017a** d_k and values d_v. Attention output is given by:

$$Attention(Q,K,V) = Sotftmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.1}$$

Where D_k is the radical(square root) of the number of patches, a process which avoids vanishing gradients. This attention function is performed in parallel for each head, and these are concatenated back and re-projected. Figure**??**, the resulting linear output depicted in figure **??** immediately before softmax layer, which produces class probability. What is learning in this process are linear projections, first local in parallel (each attention head) and then through global projection, produce final values given by:

$$MultiHead(Q,K,V) = Concat(h_1,...,h_n)W^O \tag{3.2}$$

Where $h_i = QW_i^Q, KW_i^K, VW_i^V$ or the application of attention function to the i_{th} where parameters of multi-head function are $W_i^Q \in \mathbb{R}^{d_{mdl.} \times d_k}, W_i^K \in \mathbb{R}^{d_{mdl.} \times d_k}, W_i^V \in \mathbb{R}^{d_{mdl.} \times d_v}, W^O \in \mathbb{R}^{d_{mdl.} \times hd_v}$. Note the last parameter passed is effectively a $O_t h$ where h is the number of attention layers**Vaswani2017a**, which is, in effect, a high-level recursive application where

power law applies to recursion depth.

3.2 Vision and Convolution Transformers

Here, we initially train Vision Transformers (ViT) - which were adapted as closely as possible from **Vaswani2017a**'s application within NLP by **Dosovitskiy2020** for image classification, a high-level overview can be seen in figure**??**. Changes are additional for the final MLP. **Dosovitskiy2020** also shows that attention head can be mapped onto convolutional feature maps.

Both require flattening of images into 2D exclusive patches $x \in \mathbb{R}^{H \times W \times C}$ where C is the number of channels; H, W are the input image height and width; N is patches of image resolution; $(H \times W)$ is the resolution dimension of P^2 where $W = H$. This is simply $N = \frac{W^2}{P^2}$.

The embedding process that **Dosovitskiy2020** introduces is token embedding **E** where $\mathbf{E} \in \mathbb{R}^{(P^2C) \times D}$ (where D is vector size) and positional embedding $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1)}$, which is added to patch embedding to retain position information given by:

$$Z_O = \left[\mathbf{x}_{class};\quad \mathbf{x}_p^1\mathbf{E};\quad \cdots;\quad \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos} \tag{3.3}$$

Where E_{pos} is learnable positional embedding shown in dark pink in figure **??**, and is added to the concatenated patch embedding(s) back to a final $1 \times n$D Vector, hence both global and local self attention can be learned.

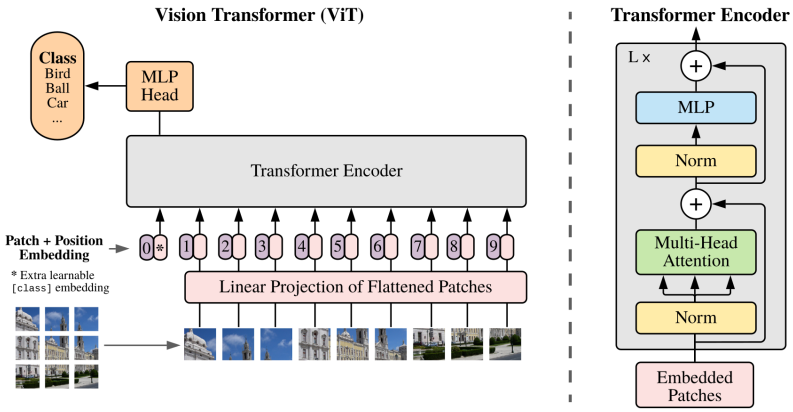


Figure 3.3: High level overview of Initial Implementation of ViT **Dosovitskiy2020**

Model	Embedding Dimension	Heads	Layers	Params	Image Resolution
ConViT-Ti	192	3	12	5M	224×224
ConViT-S	238	6	12	22M	224×224
ConViT-B	768	12	12	86M	224×224

Table 3.1: Efficient ConVits trained adapted from **Touvron2020a**

3.2.1 Training Approach

Transformers require extremely large amounts of data to train, $10\times$ typically larger than any available IAQA dataset. We combat this in three ways:

Transfer Learning leveraging transformers have been trained on ImageNet1k**Ridnik2021**;

Data-efficiency using state of the art(SOtA) data efficient transformers**DAscoli2021**, **Touvron2021a** that are able to train on smaller datasets that have been able to train on subsets of ImageNet1K;

Augmentation using SOtA**Buslaev2020a**, **Riba2020** approaches such as reflection padding;

We use pre-trained data efficient transformers developed by **Touvron2020a** models on image net, which introduces soft inductive biases (initialization patches through convolution) developed by **DAscoli2021** and keeping input dimension unchanged at $3 \times 224 \times 224$. Images are resized from the original to the longest edge. In order to preserve computational information, we train on both zero-padded and reflection padded images, where images are grey-scale 224×224 these are stacked to three equal channels as a simple means to ensure bias is not introduced to the model. We train models of three sizes, of modified data efficient transformer ConViT $\in \{Ti, S, B\}$, which are**DAscoli2021**'s adaptions of DeiT-Ti,DeiT-S,DeiT-B**Touvron2020a**.

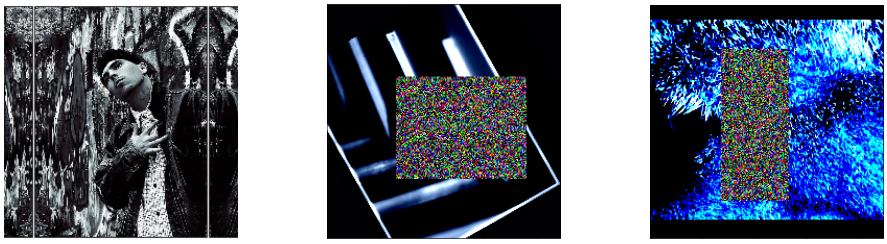
The models have a final, fully connected layer of $[1 \times 1000]$ classes, and therefore we add a fully connected layer of $[1 \times 2]$ using PyTorch `nn.Layer(1000,2)` to add a finally fully connected layer to the ConViTs model. We trained ConViT-Ti using a formative grid search, as a new hyper parameter is introduced as a modified Deit-Ti model. This requires significant search, and therefor we only perform a grid search on ConViT-Ti. Trained adjusting hyper parameters incrementally within grid searches for 10 approaches. We also train ConViTs $\in \{Ti, S, B\}$ shown inn Table **??** with no changes to hyper parameters to provide insights into how including embedded dimension, number of attention heads effects performance on test set. Finally, we train the best model on the best performing hyper parameters.

3.2.2 Augmentation

Data Augmentation is an important part of regularisation in deep learning **Kukacka2017**; it is a powerful way to mitigate against over-fitting during training and supports better generalization**Shorten2019** - ensuring that validation error decreases with training error, we implement conventional methods**Mikolajczyk2018** alongside state of the art methods proposed by**Riba2020**, **Buslaev2020a**.

Here, augmentation of images is in two stages: first, all images are square padded to zero to preserve com positional information, or reflection padded**Buslaev2020a** and resized to $3 \times 224 \times 224$. Larger image configurations are possible, however for consistency in baseline comparisons and IAQA approaches we maintain this a base dimension. During training, we augment applying random occlusion/erasure of images to remove a random patch of image for 16×16 pixels, random rotation, spatial transforms such as shear and colour-gitter, gaussian noise (5,5) filter and gaussian blur (5,5) filter in addition to square distortions.

We also use occlusion patches randomly applied to images of the same dimension of image patches, this is to strategically remove patch sized areas of the image while allowing the model to train attention on other areas of the image.



(a) Reflection Padding to 224×224 with Albumentations

(b) Training Augmentation Rotates and Random Erasure/Occlusion

(c) Training Augmentation Zero Padding, Random Erasures

Figure 3.4: Augmentation Examples

3.2.3 Data Preprocessing and Normalization

An essential part of optimising learning, both to improve training results and ensure consistency of method elsewhere. All image normalising using image change $\in \{R, G, B\}$ mean given by:

$$\mu = \frac{1}{n}\sum_{1}^{n} x_i \tag{3.4}$$

And standard deviation given by:

$$\sigma^2 = \frac{1}{n}\sum_{i}^{n}{}^2 - \mu^2 \tag{3.5}$$

Where n is number of pixels and x_i is i_{th} a pixel value of 0-255.

Resulting tensors are 32-bit floating points (FP) between 0-1, mean and standard deviation RGB values $\mu = 0.485, 0.456, 0.406$, $\sigma = 0.229, 0.224, 0.225$.

Ground Truth Discretization

AVA ground truth is derived by two processes. MOS scores are in categories $1-10$ the MOS can be computed neatly by:

$$\mu = \sum_{i=1}^{n} a_i \cdot b_i^T \tag{3.6}$$

Where a is a normalised vector of scores and b is a sequence vector $1-n$ [1]. Mean scores are then thresholded by:

$$s = \begin{cases} 1 & \geq \mu \geq 5 \leq 4\sigma \\ 0 & \leq \mu \leq 5 \leq 4\sigma \end{cases} \tag{3.7}$$

Where μ is given by eq. **??** and σ is standard deviation of overall MOS scores. This results in a small number of outliers (just 97 images out of 255508) excluded from training, test and validation sets; for many within the literature the convention is exclusion of MOS $\pm 4\sigma$.

3.2.4 Hyper Parameter Search

We perform several grid searches of hyper parameters by defining a function called within a training class that initialises existing code using Python's inbuilt sub process module, which augments rather than rebuilds the model.

The purpose of a conduction grid search is not to comprehensively tune the base model, but to assess how newly introduced hyper parameters ('locality strength' and 'locality up to layer') which incorporate convolutional inductive bias effectively adapt the model to the target domain.

This is for methodological consistency, and to enable tuning of newly introduced softer inductive bias GPSA**DAscoli2021**. The assumption made is that a degree of due consideration is made by **DAscoli2021**, **Touvron2021a**.

[1] Presented in this way for conceptual clarity and to resemble as closely as possible Python's numpy operations

Training time for 10 epochs is 25 minutes; where unique parameter combinations are used, this was a total 3 days per grid search and also involved ensuring that the model can be reinitialised. Weights were saved for every epoch of each model in the grid search.

parameters	Constraints	Intervals
locality strength	0.5-1.5	10
learning rate	0.001-3	3

Table 3.2: Formative Grid Search Parameters

3.2.5 Test Train Validation

We employ a test train split defined within the literature**Murray2012**. For convenience, **Talebi2018** have made available a .csv file which has image references alongside test, training and validation images. For the reproduction of results and methodological clarity, we have also made available .json files of test image IDs. $\in \{test, train, validation\}$ image numbers are **19928, 223795, 11779**. Figure **??** shows a bar plot of $\in \{test, train, validation\}$ sets.

3.2.6 Models

We train from/create several transformer models built on **Touvron2020a**'s DeiT, with added GPSA**DAscoli2021** and **Wu2021**, figure **??** show addition of GPSA at local level.

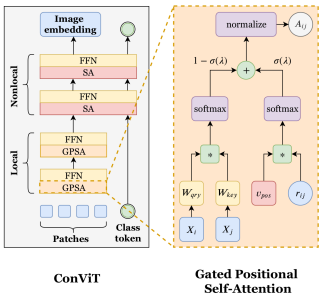


Figure 3.5: ConViT GPSA Diagram **DAscoli2021**

We train a true hybrid model CvT which maps convolutional layers onto attention patches, which does not have pre-trained weights available. The model architecture can be seen in figure **??**:

This provides metrics on a model that does not have a mechanism for switching between convolution and self attention, but incorporates *as is*.

We train ConViT as an architecture that introduces convolutions as softer biases, where architecture is a positional layer that can be initialised to allow convolution initialisation of attention heads, with a one-to-one mapping.

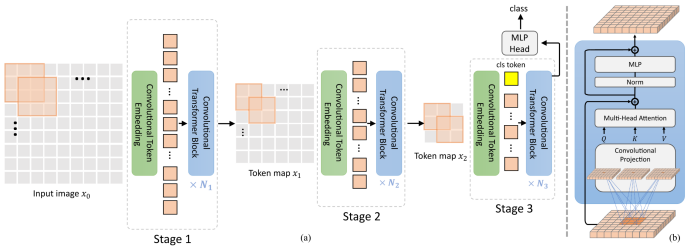
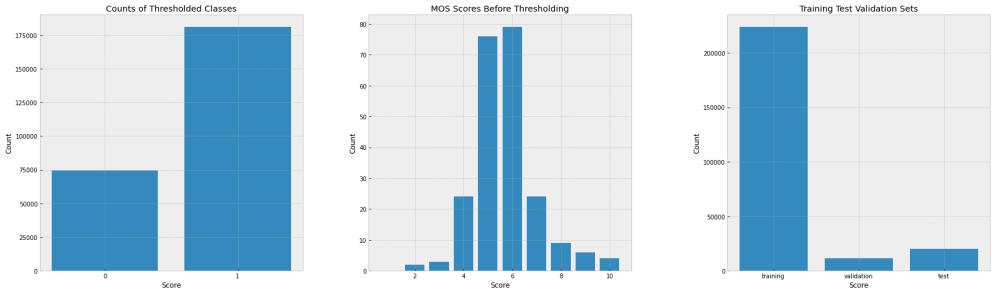


Figure 3.6: CvT Architecture and Pipeline **Wu2021**

3.2.7 Addressing Class Imbalance

When thresholded into positive and negative image classes, the AVA dataset is heavily imbalanced towards the positive class with ration of (0.43:1).

There are approaches to addressing class imbalance that involve defining a loss function and using balanced accuracy during training. However, we found that this was overly complex for a binary classification problem. Therefore, to address class imbalance during the training of ResNet $\in \{18, 50, 152\}$, we employed a data sampler to over sample the minority class during training. This further enabled repeated augmentation of the minority class which would not be possible if using a loss function. Class weights for the data-sampler are given by **??**.



(a) Binary Class (Thresholded) **(b)** MOS Scores **(c)** Training, Validation, Test

Figure 3.7: Class Imbalance and Test Training Validation Splits

3.2.8 Research Methodology Summary

We apply domain adaptation transfer learning, and train both baseline models (ResNets) to enable side-by-side comparison of ViTs, CvT's and CNNs, adding a fully connected *linear* layer which outputs a 2D vector of $n \times c$ elements, for each class corresponds to a number of classes $y = xA^T + b$ where A^T is the transposed input of the preceding layer; b is some learnable bias.

For training (batch-wise), validation (set-wise), and test (set-wise) inference, we apply a softmax function to produce class probables of the linear layer. We threshold data into binary classes from a continuous probability distribution that is calculated for each image based on a total number of votes, where each image receives an average of 210 votes.

Thresholding given by eq. **??** is standard across IAQA literature on the AVA bench marking dataset. We also perform a grid search to establish a new hyper parameter, introduced by **DAscoli2021** *locality strength*. All models are pre-trained on ImageNet1k, and we train all models for 10 epochs with all layers unfrozen (trainable).

We try to apply a range of data augmentation techniques as part of training, which is doubly important as we have employed a data sampler that oversimplifies the minority class. We also outline a case for implementing convolution to introduce a soft inductive bias to the CvT's training to see if this makes the best of both transformers and CNNs, and also show how two very different models (CNNs) and ViTs learn.

4 Results and Discussions

4.1 Experimental Settings

The settings for all experiments were conducted using test trains split defined by **Talebi2018**, and data augmentation outlined in chapter **??**, with input of raw images resized to $3 \times 224 \times 224$. To enable effective side-by-side comparison with baseline, state of the art pre-processing steps are kept the same. We also used transfer learning from the same source dataset, ImageNet1k. Transformers and CNNS are, however, inherently different - for example, learning rate alongside approaches such as warming-up**Zhang2021** are necessary for transformers to converge, therefore some aspects - such as learning rate - change during training of baseline and transformer models.

4.1.1 ResNets and ConViTS

We conduct experiments to obtain baselines using ResNets $\in \{18, 50, 152\}$ and ConViT models $\in \{tiny, small, base\}$, which are pre-trained on ImageNet-1k, and added an additional, fully connected layer to correspond to binary classes. Hyper parameter selection is a crucial part of tuning and adapting models to the target domain.

Table 4.1: Hyper Params

Hyper Param	ResNet innitial	ConViT	Equal Conditions
Learning Rate	0.0001	0.0005	0.0003
Colour Jitter	0.5	0.5	0.5
Optimiser	*Adam*	*AdamW*	*AdamW*
Batch size	50-250	10-200	10-200
Loss Function	*crossentropy*	*crossentropy*	*crossentropy*
Cool down epoch	-	5	0
Scheduler	*StepLR*	*cosine*	*cosine*
Opt. ε	-	1e-08	
Local up to layer	-	10	-
Locality strength	-	4.0	-
lr noise pct.	-	0.67	0
Warm up epochs	-	5	3
Warm up lr	-	1e-06	1e-06

4.1.2 Optimiser and Scheduler

These differ slightly between ConViT and baseline models.

For ConViTs We use AdamW with a Cosine[1] learning rate Scheduler;

[1] https://github.com/rwightman/pytorch-image-models/blob/master/timm/scheduler/cosine_lr.py

for ResNets we use Adam and AdamW with Stepped learning rate scheduler and cosine scheduler;

for ViT and CvT we use Adam with Cosine learning rate scheduler;

We use an Adam optimiser**Kingma2015** as gradient base optimisation for simplicity of implementation and computational efficiency. We use a Cosine Scheduler by **DAscoli2021** for ConViTs, as there is clearly a trade-off in keeping domain parameters the same. Both schedulers are a form of simulated annealing, which allowed for a change in learning rate over time - mitigating against the challenges of gradient based approaches, such as saddle points.

We train networks in separate code environments[2], using code supplied by **DAscoli2021** for ConViTs and CvT. We train all models under equal conditions, in the same code environment, with identical training and data augmentation pipelines. Various aspects, such as data augmentation, are inherent features of training pipeline.

4.1.3 Loss Function Vs Data Sampler

Loss function is necessary to minimise the difference between GT labels, and where minimisation of loss or error between GT and predicted result is required. While formative experiments were conducted with focal loss, we adopt a cross entropy loss function. While **Lin2020** have shown that focal loss, combined with cross entropy, have been shown to be effective techniques in addressing class imbalance.

We use cross entropy with a sampler, where random sampling from the minority class with a weighted probability is used to address this. This is effective, as it allows repeated augmentation of the minority class during training. Further, this is the technique used in code for training on ImageNet-1k by **Touvron2020a**, **DAscoli2021**, and therefore provides methodological consistency across baseline and vision transformer models .torch.utils .data.sampler.WeightedRandomSampler

Where samples are computed for every batch given by:

$$Sc = \frac{\sum_{i=1}^{Bn}}{\sum_{i=1}^{Cn}} \tag{4.1}$$

Where Sc s sample ratio Bn is the number of batch samples, and Cn n is the class count of each batch.

4.1.4 Software and Hardware

The experiments were conducted using a single Nvidia Tesla P100 12Gb GPU, with CUDA 11.2. The software used for training was PyTorch open source framework, with data am-

[2] https://github.com/facebookresearch/convit

Software	Version	Purpose
torch	1.7.0	Training ConViTs
timm	0.3.2	Building ConViTs
albumentations	0.4.6	Data augmentation
torchvision	0.8.1	Pre-trained models

Table 4.2: Software

munition using both Torchvision's native fictions alongside Albumentations**Buslaev2020a**. ConViT models were built using timm[3]. All software is implemented using Python 3.7.

4.2 Quantitative Results

4.2.1 Training

ResNets

Training ResNets for 10 epochs shows differences in overall accuracy. These are shown in figure **??**. In training and validation accuracy, loss, and balanced accuracy, all layers were unfrozen to allow pre-trained models to be fully adapted during training.

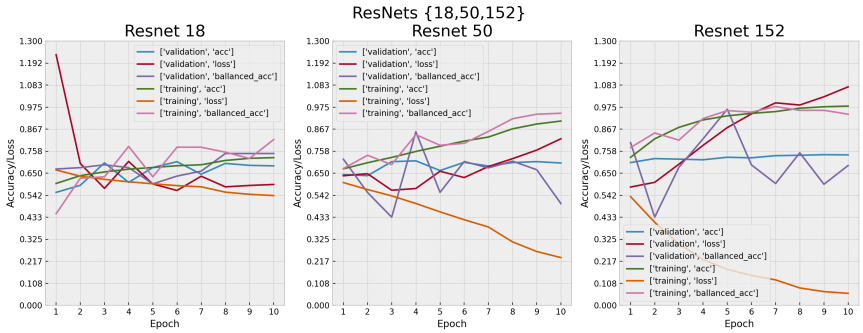


Figure 4.1: Side by Side Comparison of ResNets

Note that only ResNet18 converges after pre-training, which importantly shows that ResNet18's depth may not have enough trainable parameters or be deep enough to produce the best results. Further, note that after convergence there is very little improvement of training or validation accuracy.

Balanced accuracy shows more instability, indicating that accuracy in itself may be 'masking' model fitting. However, both appear similar where equal numbers of positive and negative classes are sampled in each training batch.

ResNet152 continues improvement in accuracy and balanced accuracy (there is also less difference). This indicates that when a model has sufficiently trainable parameters and

[3] https://github.com/rwightman/pytorch-image-models/tree/master/timm

is deep enough that there is a high degree of similarity to features that are trained on ImageNet21k.

	Resnet 18		Resnet 50		Resnet 152	
	Validation	Training	Validation	Training	Validation	Training
Accuracy	0.706	0.726	0.710	0.906	0.741	0.980
Balanced Acc	0.747	0.815	0.854	0.944	0.964	0.977

Table 4.3: ArgMax Resnet Training Metrics

Table **??** shows the differences in training accuracy and balanced accuracy - there is a clear improvement in increasing model size over the same number of epochs.

CvT and ConVits

We train both CvT**Wu2021** and ConViTs to enable side by side comparison. Results from training CvT**Wu2021** show that a model will converge after 5 of epochs (see figure **??** right), however, the model quickly plateaus. This can be seen in figure **??** left.

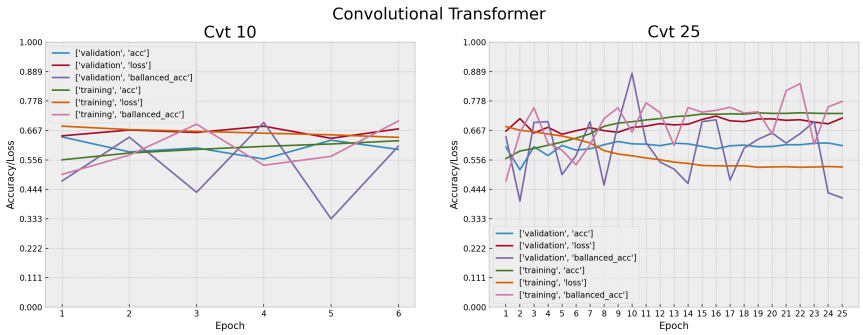


Figure 4.2: Hybrid Convolutional Transformer Training

Further, it can be seen that training inference is unstable (both training and validation (balanced accuracy) vary significantly), which is consistent with observations elsewhere - often requiring a warm-up (tapering of learning rate) during initialisation. Further, given the inclusion of a data sampler, balanced accuracy does not provide a clear prediction of model accuracy and therefore subsequently trained models do not report this. **??** left.

Comparing results between CvTs and ConViTs, both without pre-training on ImageNet1k, shows erratic training performance on validation sets.

A marked improvement can be seen in the use of pre-trained models with a 5 epoch warm up, shown in **??**, where it is also clear the model size has a bearing on validation accuracy (although this appears to be slight).

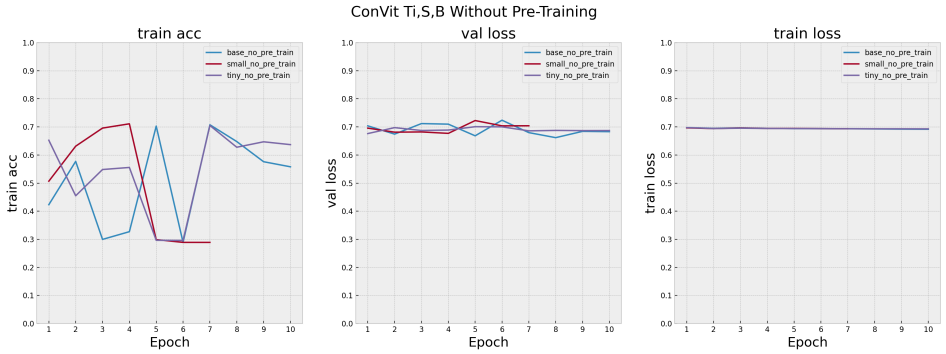


Figure 4.3: ConViT Without Pre-Trainign on ImageNet1k

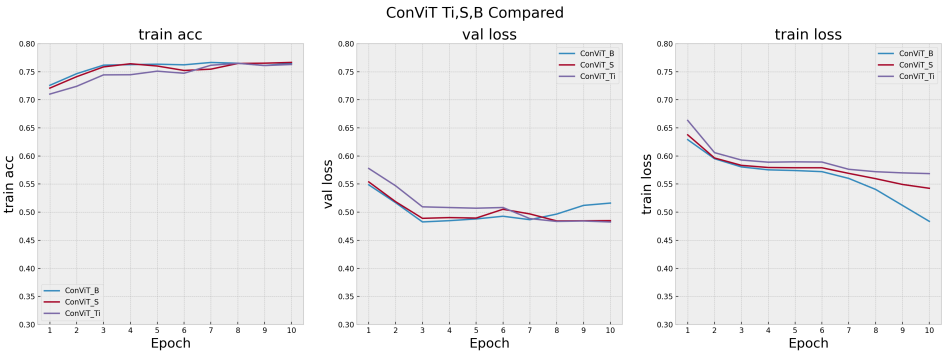


Figure 4.4: ConViTs With Pre-training

Hyper-Parameter Tuning and Data Augmentation

Conducting a grid-search over learning rate and locality strength, it is clear that learn locality strength has a slight effect on model performance. Training was conducted with no warm start epochs. This shows that decreasing learning rate improves overall performance. This is shown in figure **??** left. The newly introduced hyper-parameter has some effect on overall performance.

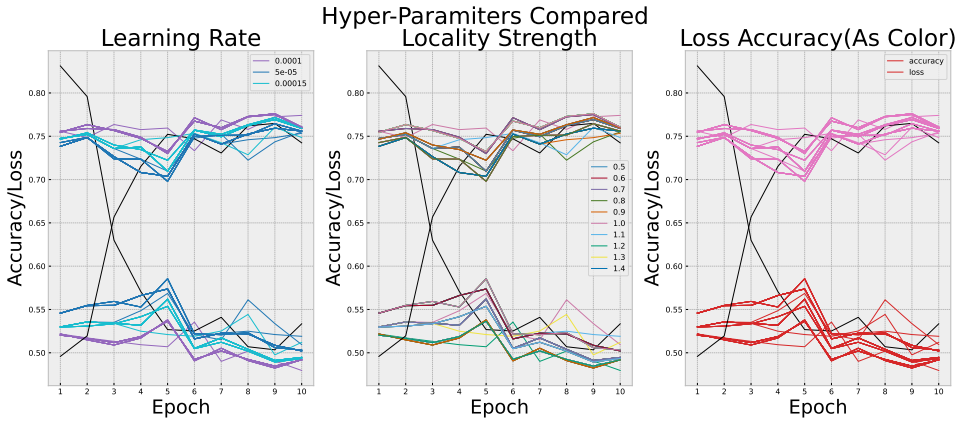


Figure 4.5: ConViT Ti Grid Search Between Learning Rate and Locality Strength (With Control)

Control training, including warm up epochs, was introduced to demonstrate difference

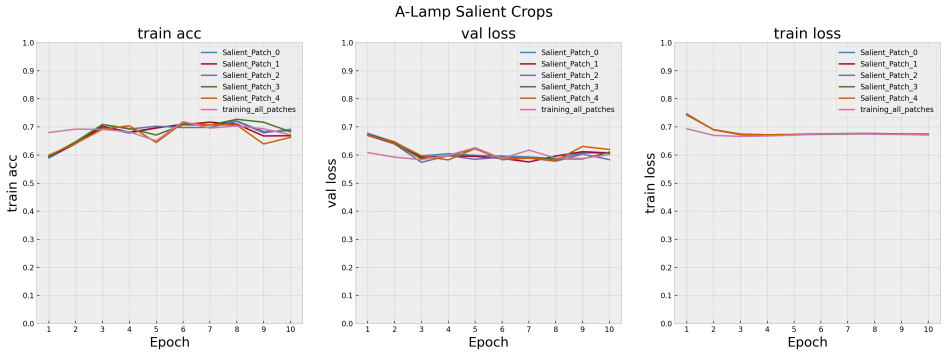


Figure 4.6: Comparison of ConViT Ti Trained on Salient Patches from A-Lamp Model **Ma2017**

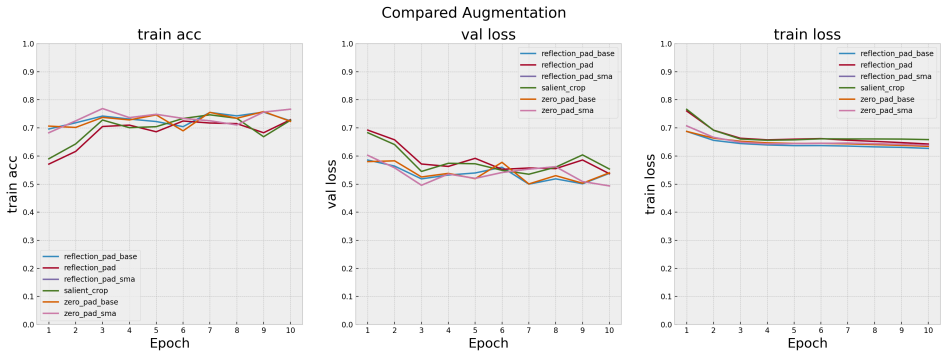


Figure 4.7: Comparison of Augmentation on ConViT B and S

between training with and without warm starting. This is shown in black on figure **??**. This also shows that using pre-training that not using a warm start results in relatively high accuracy, with almost no training on the AVA dataset.

In order to compare spatial data augmentation, we performed training of salient crops where $5 \times (224 \times 224)$ sub patches of each image were taken using the attention algorithm provided by**Ma2017**. This was in order to gauge how a self attention model might train on data, and whether higher resolution patches might enhance training. Training on all five sub patches separately, alongside training on all patches pooled, is shown in figure **??**.

Combining all demonstrates best performance, however all perform less well than a global/non cropped image, which indicates that the ConViTs are learning composition information.

We also compared training on zero-padding to square and reflection padding. The logic of this is that reflection padding maintains composition data, while reflecting the shortest edge effectively also encodes an image border.

This shows some improvement during training by reflection padding.

	Starting State		Augmentation			Search
	Pre-train	No Pre-train	Reflection Pad	Zero Pad	Salient Crop	Grid
ConViT B	0.766	0.708	0.756	0.758	-	-
ConViT S	0.767	0.711	0.769	0.769	-	-
ConViT Ti	0.765	0.705	0.749	0.714	0.747	0.776

Table 4.4: ArgMax Validation Results after 10 Epochs

4.2.2 Test Metrics

Table **??** shows all results evaluated on AVA test data: 19k images on reflection padded and normalised images. Note that accuracy does not improve monotonically with model size for ConViTs, but does for ResNet models - however, F1 score does appear to obey a monotonic relationship with size of ConViT Model.

We also compared this with zero-padded images to square, as this might reflect how images are processed in the wild:

Here, each ConViT out performs in accuracy, however it is interesting to note ResNet152 out performance on balanced accuracy and F1 Score ConViT B.

	Accuracy↓	Balanced Acc.	F1	N epochs	Weight Save Arg
ViT B	39.96	26.66	39.93	10	Last Epoch
CvT	63.90	76.68	52.47	25	Max Acc. Epoch
ResNet 18 R	70.23	58.08	65.66	10	Last Epoch
ResNet 50	70.85	99.54	41.70	10	Last Epoch
ResNet 18	71.31	86.84	54.74	10	Last Epoch
ResNet 152	72.69	87.66	55.84	10	Last Epoch
ConViT Ti	73.17	95.52	49.58	10	Last Epoch
ResNet 50 R	74.13	72.06	65.69	10	Last Epoch
ConViT Ti	74.33	90.90	55.42	10	Max Val. Acc. Epoch
ConViT B	74.58	85.53	59.65	10	Last Epoch
ConViT S	75.02	87.58	58.85	10	Last Epoch
ConViT S	75.06	90.22	56.94	10	Max Val. Acc. Epoch
ResNet 152 R	75.63	60.32	70.82	10	Max Val. Acc. Epoch
ConViT Ti R	76.29	72.33	68.07	10	Max Val. Acc. Epoch
ConViT S R	76.29	72.33	68.07	10	Max Val. Acc. Epoch
ConViT B	76.47	81.00	64.59	10	Max Val. Acc. Epoch
ConViT B R	76.82	70.38	69.35	10	Max Val. Acc. Epoch
NIMA	77.25	69.51	70.56	0	Pre-Trained
ConViT B R	**78.11**	**73.17**	69.92	25	Max Val. Acc. Epoch
MLSP	81.65	66.68	76.08	0	Pre-Trained

Table 4.5: Evaluation on Refection Padded Normalised Images Square Padded on 19k AVA Test Set

We show that, when trained on best hyper parameters (from the grid search conducted [learning rate <0.0015 and locality strength 1.5]), ConViT B outperforms NIMA**Talebi2018**, whose results we reproduce on their test set with a pre-trained model made available alongside their publication. Both the F1 score and balanced accuracy of MLSP**Hosu2019** are lower than ours. This is in spite of the potential that training an end to end binary classifier is more challenging, with a smaller, fully connected final layer output of 2 rather than 10 (one for each MOS probability produced by normalised MOS score over 10 bins).

One further important feature to consider is that **Talebi2018**, **Hosu2019** both sample from much larger initial images, and we only train on 224×224 resized images. We also show a confusion matrix (each column represents the entirety of the 19k AVA Test set sample space). Here, we see that ConViT B outperforms in maintaining a high number of TPs while not sacrificing the the negative minority class TN.

	ResNet 50 R	ResNet 152 R	ResNet 18 10	ConViT Ti R	ConViT B R	ConViT S M	ViT B	MLSP	NIMA
tp	0.62	0.58	0.66	0.64	0.63	0.70	0.52	0.65	0.62
tn	0.12	0.18	0.05	0.13	0.14	0.05	0.08	0.17	0.15
fp	0.09	0.13	0.05	0.08	0.08	0.01	0.19	0.06	0.08
fn	0.16	0.11	0.23	0.16	0.15	0.24	0.20	0.12	0.15

Table 4.6: Confusion Metrics on 19k AVA Test-Set 19k

Table **??** shows nomenclature used in tables (**??**, **??**), as models performed differently according to whether zero-padding or reflection padding were used during data augmentation. We also show type number of layers and dimension of resized images.

	Augment	Layers	Type	Dimensions
ResNet 18 Z	Zero Pad	18 Residual	CNN	224×244
ResNet 50 Z	Zero Pad	50 Residual	CNN	224×244
ResNet 152 Z	Zero Pad	152 Residual	CNN	224×244
ConViT Ti Z	Zero Pad	12×3	Conv. Vision Transformer	224×244
ConViT S Z	Zero Pad	12×6	Conv. Vision Transformer	224×244
ConViT B Z	Zero Pad	12×12	Conv. Vision Transformer	224×244
ResNet 18 R	Reflect Pad	18 Residual	CNN	224×244
ResNet 50 R	Reflect Pad	50 Residual	CNN	224×244
ResNet 152 R	Reflect Pad	152 Residual	CNN	224×244
ConViT Ti R	Reflect Pad	12×3	Conv. Vision Transformer	224×244
ConViT S R	Reflect Pad	12×6	Conv. Vision Transformer	224×244
ConViT B R	Reflect Pad	12×12	Conv. Vision Transformer	224×244
ViT B R	Reflect Pad	12×12	Vision Transformer	224×244
CvT R	Reflect Pad	3×10	Conv. Vision Transformer	224×244

Table 4.7: Number Parameters and Training Time Per Epoch 230k Images on 1 GPU-batch size 10 Inference on 19k Images on AVA dataset

4.3 Exclusive Set Analysis

This section shows images from the subsets shown in Table **??**. These images are a complement set of the union of all other model inferences on the 19k image AVA test set, $M = \overline{\cup}_{i=1}^{n} F_i$ where F is all other models. This is performed for each model and each inference type $\in \{tp, tn, fp, fn\}$. The results are shown in table **??**, we compute this for the best performing model of each type. This shows as a discreet value, the number of images where a single model over or under performs against all the others in the entirety of the 19k.

	tn	fp	fn	tp
ResNet 50 R	131	17	408	7
ResNet 152 R	343	2	599	1
ResNet 18 10	52	122	286	13
ConViT Ti R	86	3	170	1
ConViT B R	129	9	201	4
ConViT S R	5	76	9	24
ViT B	341	157	2422	18

Table 4.8: Number of Unique Images by Confusion Metric Categories

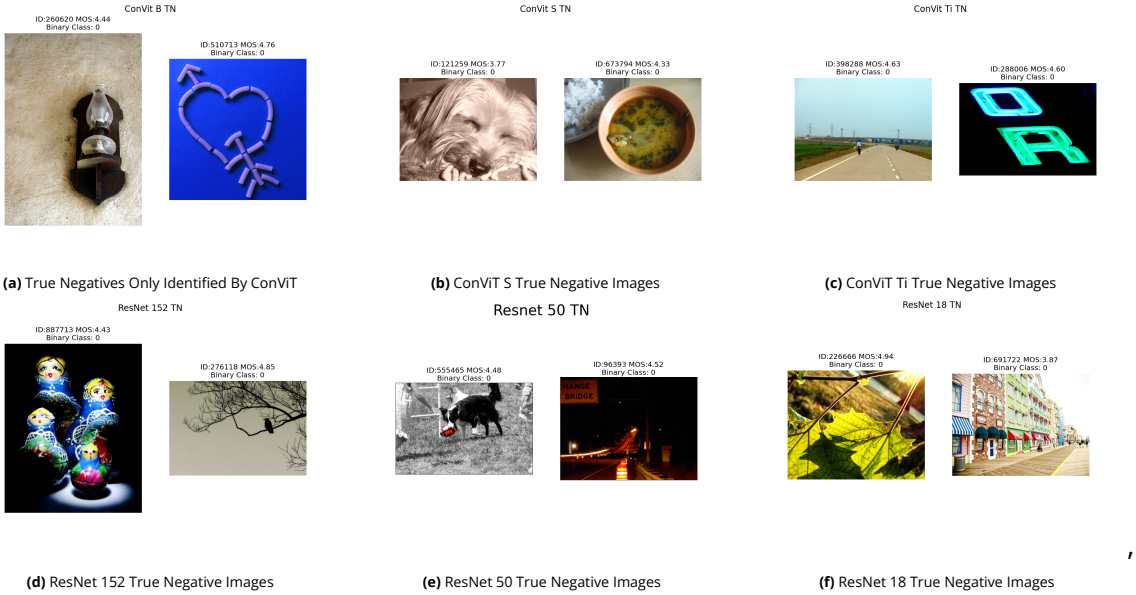
An interesting observation is that ViT, despite performing poorly overall, still correctly identifies a significant number of images that are challenging for other models. Overall, ResNets have far more false negatives where they alone under-performed. This is an important observation, as this is for the minority class. Here, one might consider whether this gap would grow with the availability of more data.

4.4 Qualitative Results

We show images from the test 19k test set - which are identified by each network in $\in \{tp, tn, fp, fn\}$ - and are members of the exclusive sets shown in table **??**. This may provide insight into the types of images that each network is performing best and worst (that is unique to that each networks inference). Images are chosen at random from each exclusive set. This is to provide insight into how each network is selecting attributes that may be visible to the human eye. It is clear from the MOS score that these images are generally from the ambiguous range of MOS.

4.4.1 ResNets and ConViTs True Negative

ConViTs (**??**, **??**, **??**) appear to pay more attention to composition information, where as ResNets - shown in figures (**??**, **??**, **??**)- appear to predict better where attributes such as good lighting and colour harmony are distinguishing features for images in the ambiguous range.



(a) True Negatives Only Identified By ConViT

(b) ConViT S True Negative Images

(c) ConViT Ti True Negative Images

(d) ResNet 152 True Negative Images

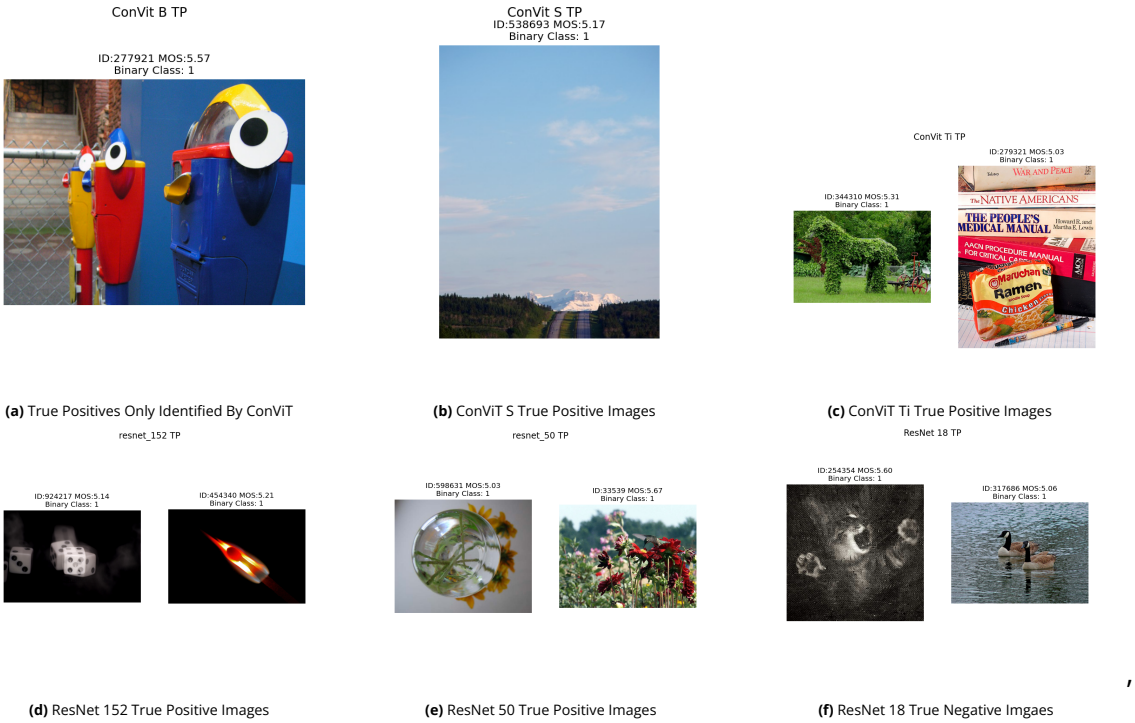
(e) ResNet 50 True Negative Images

(f) ResNet 18 True Negative Images

Figure 4.8: Examples of Exclusive True Negatives Predictions by Each Network (Exclusive Set) on AVA Test Set

4.4.2 ResNets and ConViTs True Positive Images

ConViT models appear to be less able to resolve positive class images that are not identified elsewhere. Again, it seems clear that interplay between composition and salient objects/distinguishing features, where figures (**??**, **??**,**??**) appear to show some semantic class ambiguity **??**. Left shows an animal made of grass/leaves and **??** shows an image of parking meters that have eyes.



(a) True Positives Only Identified By ConViT

(b) ConViT S True Positive Images

(c) ConViT Ti True Positive Images

(d) ResNet 152 True Positive Images

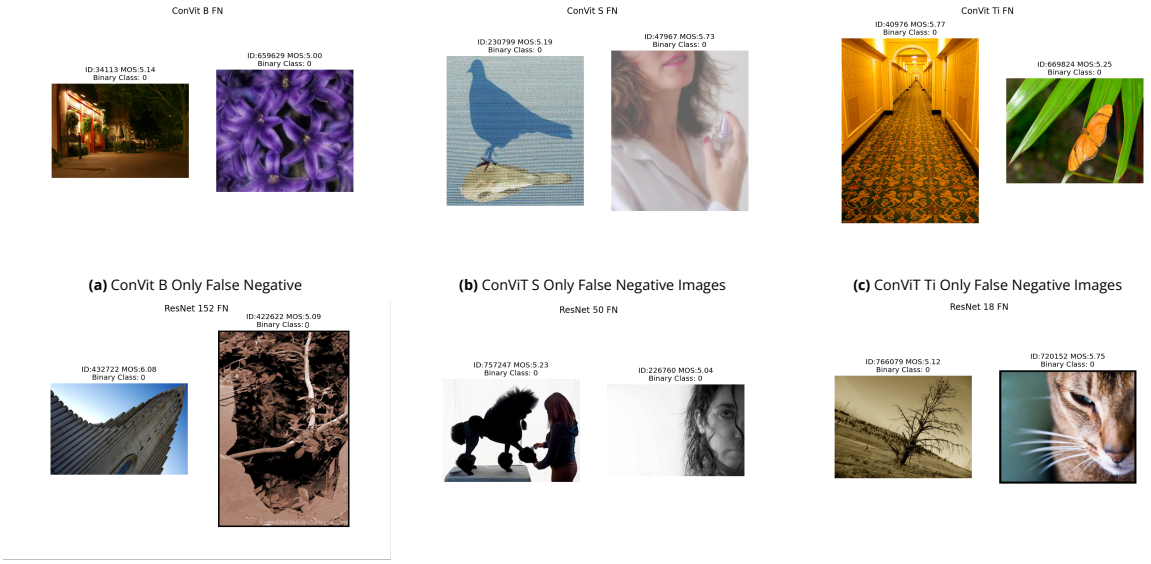
(e) ResNet 50 True Positive Images

(f) ResNet 18 True Negative Imgaes

Figure 4.9: Examples of Exclusive True Positive Predictions by Each Network (Exclusive Set) on AVA Test Set

4.4.3 ResNets and ConViTs False Negative Images

ConViT FNs are shown in figures (**??**, **??**,**??**) and appear to falsely rate images as negative with a bias again towards high level semantic information, in contrast with ResNets' FNs shown in figures(**??**, **??**, **??**).



(a) ConVit B Only False Negative

(b) ConViT S Only False Negative Images

(c) ConViT Ti Only False Negative Images

(d) ResNet 152 Only False Negative Images

(e) ResNet 50 Only False Negative Images

(f) ResNet 18 Only False Negative Imgaes

Figure 4.10: Examples of Exclusive False Negative Predictions by Each Network (Exclusive Set) on AVA Test Set

4.4.4 ResNets and ConViTs False Positive Images

(a) ConViT B Only False Positive

(b) ConViT S Only False Positive

(c) ConViT Ti Only False Positive

(d) ResNet 152 Only False Positives

(e) ResNet 50 Only False Positive

(f) ResNet 18 Only False Positive

Figure 4.11: Examples of Exclusive False Positive Predictions by Each Network (Exclusive Set) on AVA Test Set

ConViT FP images appear to show a bias for global form, shown in figures (**??**, **??**,**??**), however, not the MOS of figure **??**, which is exacly 5.0 where negative class is ≤ 5. Figures (**??**, **??**, **??**) show FPs exclusive to respective ResNets.

4.4.5 Training Metrics

When shown in the same plot (figure **??**), it is clear that ResNet50 and 152 overfit during training and that ConViTs (as well as CvT) consistently improve with accuracy in a more linear fashion. Further, it is noteworthy that ViT (pure transformer) shows a dip in initial training accuracy during training, and validation loss steadily increases.

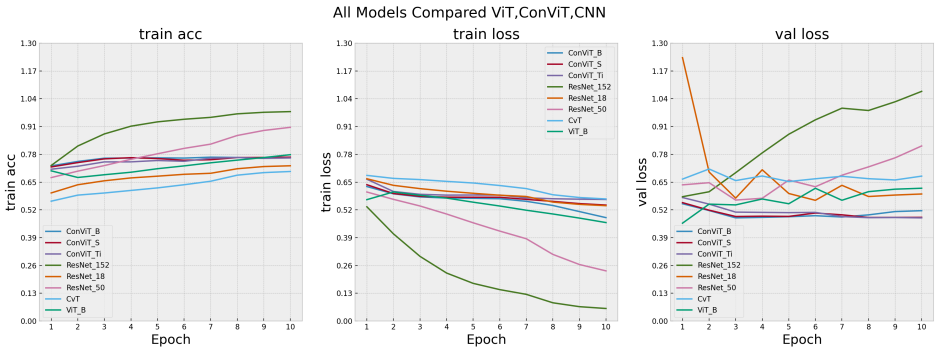


Figure 4.12: All Models Compared

High maximum validation loss for all of the ResNet models is shown in table **??**. It is also noteworthy that ViT B performs well on the validation set (outstripping all other models which declined after first epoch).

	Train. Acc.	Train. Loss	Val. Acc.	Val. Loss
ViT B	0.778	0.604	0.786	0.621
ConViT B	0.766	0.629	0.781	0.549
ConViT S	0.767	0.638	0.746	0.554
ConViT Ti	0.765	0.663	0.762	0.578
ResNet 152	0.980	0.535	0.741	1.074
ResNet 50	0.906	0.604	0.710	0.819
ResNet 18	0.726	0.665	0.706	1.232
CvT	0.700	0.681	0.625	0.711

Table 4.9: Overall Maximum Training Results Compared

We also show inference times and training times per epoch to give additional compute resources, and show how models might perform when embedded in a real world application (table **??**). ConVit B is clearly resource intensive.

	# Million Trainable Params.	Train Time Per Epoch 230k Images	Inference Time 19k Images
ResNet 18	11.2	56 mins.	3 mins.
ResNet 50	23.5	73 mins.	4 mins.
ResNet 152	58.1	121 mins.	5 mins.
ConViT Ti	5.5	78 mins.	3 mins.
ConViT S	27.3	92 mins.	4 mins.
ConViT B	85.8	144 mins.	5 mins.
ViT B	85.8	121 mins.	5 mins.
CvT	17.6	94 mins.	4 mins.

Table 4.10: Number Parameters and Training Time Per Epoch 230k Images on 1 GPU-batch size 10 Inference on 19k Images on AVA dataset

5 Conclusion and Future Work

This section outlines the contributions that we have made to IAQA as well a outlining the limitations and drawbacks in section **??**. Section **??** gives an outline of both potential for real world applications and areas of research that might provide new learning such as multi modal transformers(section **??**) Social network (section **??**) and continuous learning (section **??**). Finally we address potential applications of IAQA research and trained models in section **??**.

5.1 Main Contributions

We have shown that transformers, even when very large and could not be trained normally on datasets as small as those available for IAQA, do perform well. Further, while initial training of ConViTs on ImageNet1k is often ongoing for almost a week, that using distributed training these models easily transfer to IAQA domain. Even after only a few epochs, they appear to show high training accuracy. This makes them suitable and efficient models (when pre-trained) and further shows that there is a high degree of similarity between what has been learned on ImagNet1k and the IAQA domain, which may seem counter intuitive. This is a useful finding in itself, and is further underscored by the high performance of ViT in early epochs.

Many of the images in the AVA benchmarking dataset are ambiguous, and transformers appear to handle these more effectively - with the ability to correctly predict ambiguously scored images where the semantic content is also ambiguous.

Newly introduced hyper parameters, such as GPSA, do not appear to show improved ability to adapt. While they converge on less data, they are still not suitable for use on a dataset as small as AVA.

We showed that training does not necessarily require warm up epochs for a binary classifier for ConViTs, however this does not seem to be the case for ViT where the initial epoch shows the highest validation accuracy (both overall when compared to other models and to subsequent epochs). There are two potential ways to interpret this: first, that the soft inductive bias cushions the adaptation of pre-trained models with its soft inductive bias, and second, that the warm up epochs simply were not sufficient (set to 3 for all models). This second possibility would be verifiable by conducting more experiments and adjusting the warm up epochs of the scheduler.

The former might be verifiable by introducing convolution layers and gradually mixing heads that have convolutions with heads that do not have convolutions.

5.2 Limitations and Critical Evaluation

There are several drawbacks to the approach that was taken here. These fall into the following categories:

- Training Approach;
- Models; Experimental design;
- Evaluation.

.

We trained initial ConViTs**DAscoli2021** using available code at a high level. This made it possible to introduce hyper parameter 'locality strength' as this was a high level pragmatic feature, however we did not initially train under identical conditions, with controls such as random state, using the same scheduler and optimiser. This made the initial comparison difficult, and required retraining all models under identical conditions.

We did not make additions to models apart form adding a linear layer with outputs corresponding to a binary class, and as such are not able to assess whether adjusting the model architecture would have improved the model performance. This, however, is curtailed by the limitations of using pre-trained models where training transformers from scratch is not possible on the datasets provided or with the compute resources available to us.

While some aspects of the architecture, such as the depth and size of the model, are compered, we did not conduct an ablation study which could have provided insights into which part of network are most critical for inference and to explainability of results. This was in part as the size of models required lengthy training, which would have been prohibitive within the timescales for this project.

5.3 Future Perspectives

There is clearly a great deal of potential for more research both in the application domain of IAQA, and within image classification more widely. Where the focus of the general classification tasks on ImageNet1k is geared towards improving model performance and domain applications are focused on sub, or application domains, we make recommendations both in terms of research within the field of IAQA and applications of IAQA to industry.

5.3.1 Further Research

While we have trained a significant number of different networks to provide a side-by-side comparison, and have made adaptations to a binary classification problem, it would be interesting to see how network architectures might be adapted or how ensemble training might leverage the best of both the ViT and CNN worlds globally as well as locally.

Consideration is made here to:

- Model Architecture and spatial adjustment including Siamese approaches and unsupervised; Training as 10 class probability distribution and or regression;
- Exploring other areas of data including data dictionary, metadata and textural information.

Vision transformers offer a new perspective, both in qualitative results and in their potential to reach higher accuracy - however, CNN approaches that have inbuilt attention mechanisms still outperform out of the box ViTs and ConViTs.

Another option would be training as a 10 bin (probability distribution) to produce a more detailed metric, which might have provided a finer grained analysis as has been provided by **Zhang2021d**. This would further enable side-by-side analysis of accuracy 10 bin probability distribution, where accuracy might be compared with both the majority class as well as mean score. This would allow the comparison of predicted probability distributions alongside actual probability distributions.

Further experiments could also examine the possibility of adjusting convolutional features to adjust levels of inductive bias **??**. This might include adjusting the stride, but also looking at max or average pooling of convolution feature maps that are larger than ViT patch dimension.

Further, many of the images in the AVA benchmarinkg dataset are much larger than 224×224, therefore data is clearly lost in downsampling images. A large feature of the AVA dataset is the size of the images; a significant portion of the 64gb uncompressed dataset is effectively lost. This is further compounded by the fact that image resolution is itself a potential aesthetic attribute, and photographers may have purposefully chosen high resolution images or vice versa as a part of their competition submission.

This might require adjusting patch sizes to optimise for the task, as this may be the bottleneck in learning should larger images of $3 \times 256 \times 256$ for example.

Multi-Modal Transformers

Given that is is well recognised that the semantic content of images has a bearing on aesthetic quality**Simond2015**, **Mullennix2020**, it would be useful to examine multi-modal approaches to using both natural language and vision transformers, particularly given the successes of natural language transformers such as BERT**Devlin2019**, **Wolf2020a**. Some key areas within this are:

1. **Key areas within this are**.
   (a) A mapping of Tokeniziton and feature mapping between NLP tokens;
   (b) Training NLP for both sentiment and semantic content form image pages to enhance training and develop context aware feature maps;
   (c) Examining the nature of figurative language and the content of the image (it is well known that parts of speech and written natural language that remain challenging are to do with nuance interpretation, for example, the use of sar-

casm or irony. Properties that are intuitive parts of human social interaction;

(d) Examining what further content there is available on a dataset that is not include with AVA - each individual image page on DPChallenge.com has a great deal of momentary from DPChallenge.com members.

The text descriptions of each image provided by the users of DPChallenge are rich examples of human annotation.

> The colors and the smallest structure of bugs is amazing. You have brought to ones eye all we don't see.**Brendel2021**

A part of this may also be mapping an online social network (DP.Challenge.com)[1], something that in itself might prove to be fascinating research.

This approach might also yield explainable insights into the transformers beyond attention**Chefer2020**, with the training of NLPs providing contextual information that might be purposefully excluded or altered.

Continuous Learning

Continuous learning is recognised as being of particular importance where context is vital **Chen2016**, **Lomonaco2020**, and IAQA is one such area where context and new available data take precedence.

For instance, DPChallenge has added 120k new images since AVA Bench marking dataset was scraped and compiled in 2012. We have web-scraped these along with the associated meta data. With more than one new challenge being created per calendar month, which, in addition, have frequently updated comments.

One particularly exciting area within this is the potential for training and predicting image quality in real time, and comparing with completion ground truth as this develops. This would also have the potential to be leveraged for commercial purposes, where finding and ranking new image content may be part of obtaining a competitive advantage - especially where photo and advertising need new content that is of high quality that also reflects, for example, current fashion trends. Such a model might be able to provide insights into what is different between trending styles and well established image quality features.

IAQA is a rich field, with many publications - <300 associated with the AVA benchmarking dataset. Performing a meta analysis of the domain, including mapping citation relationships and datasets, would be a rich exercise and contribute new knowledge to the field. This would also support the production of a data dictionary for IAQA.

The code repository associated with this dissertation provides a formative example of how this might be structured[2]. Many of the IAQA datasets reviewed in chapter **??** shown

[1] https://www.dpchallenge.com/
[2] https://github.com/fdsig/iaqa

in table **??** and in the appendix **??**. These provide subclass granularity, as well as quality classes.

Social Networks

The online community within dp.challenge.com is clearly well established, with trust and supportive feedback being provided by members. In addition to leveraging the comments data outlined above, it may also be a fruitful avenue of research to use deep learning for community learning and to map the online community **Jin2017**, **Wu2020b**. This may also feed into other areas of research where online community is central, such as deep fake detection **Ajao2019**.

The dp.challenge.com community might also have a vested interest in such areas, with a growing potential for deep fake images being presented in competitions. Further, this may also have the benefit of providing new meta data on the AVA benchmarking dataset, with the ability to leverage data on voting patterns within competitions - something that is presently absent and has been the source of criticism of the AVA benchmarking dataset.

5.3.2 Recommendations and Applications

While IAQA as a classification task in its own right is interesting, there are also several areas of commercially that could be further developed - such as a mobile application. This might involve research into how best to prune networks and make different trade-offs, such as whether false positives or false negatives are preferable. Further, model size would be an important consideration within on device real time IAQA inference - increasingly compact and efficient CNNs**Feng2019** might be appropriate. Many of the models trained would be too large to use for inference on a mobile phone device.

Areas that haven't been explored are:

- Art applications have not been fully explored;
- Predicting individuals as well as groups as a commercial application;
- Embedding quality estimation within a mobile application.

Novel applications may come from *deeper* rather than *broader* research. For instance, using art databases to learn the aesthetics of particular periods of art history. This would involve formulating IAQA problems in different ways, such as conducting unsupervised learning to, for example, learn the aesthetic space of a particular artistic movement.

Aesthetic classifiers as filers on databases: this may be useful in areas such as medical imaging, to be able to either apply or select from high quality images alongside having applications in mobile device filter.

On-device image enhancement: in a world where much of our lives are recorded on mobile devices, this is clearly something that has a great deal of commercial potential, as is automated enhancement or selection of product photographs such as **idealo2021**

A Appendix

A.0.1 IAQA Metrics Used Where Models are Trained as Regression Problem

The approach to IAQA on AVA has generally been of binary classification however a number of approaches treat IAQA as a 10 class problem reproducing a a probability distribution via softmax. Where this has been the case a predicted MOS has been computed as in eq. **??** and mean correlation metrics have been used alongside classification metrics to evaluate model performance:

Spearman's Rank Correlation Coefficient

Spearman's rank is use to evaluate model performance by **Talebi2018** on the MLSP model. This is used as an alternative to Pearson's linear correlation (PLCC) and is a measure of a monotonic relationship rather than linear.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{A.1}$$

Where d_i is the difference in ranks of predicted and actually mean IAQA scores.

MOS (Mean Observed Score) Ground Truth (GT) and predicted value are given by softmax mapping of the final layer of the network given by:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{A.2}$$

i and j are respective output values of the network:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \quad for\, i = 1,2,\ldots,n \tag{A.3}$$

Where ground truth and predicted MOS values of images are given by vectors.

$$y_{predicted} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \tag{A.4}$$

$$y_{ground} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \tag{A.5}$$

Pearson Linear Correlation Coefficient (PLCC)

This gives conventional correlation between variables, and is a measure of linear relationships. The r score is given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{A.6}$$

Where x is the predicted score and y the ground truth score and $\overline{x}, \overline{y}$ are respective means (MOS score).

A.0.2 Single Attention Head of ConVit Models

Figure **??** shows a high level overview of a single attention head used in all ConVit Architectures.

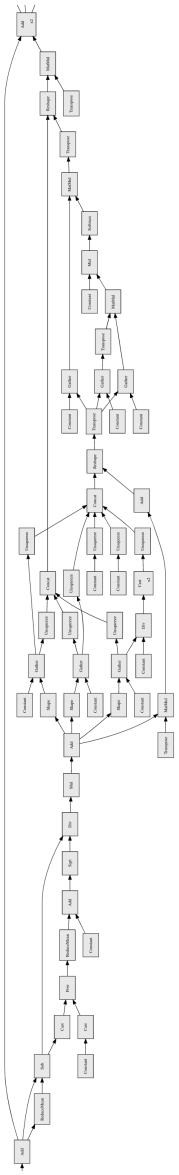


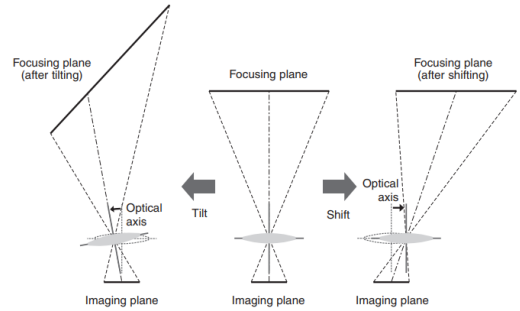
Figure A.1: ConViT Layer

A.0.3 Grid Search Function

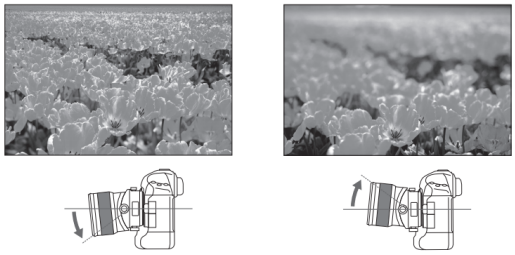
```
def grid(self):
        para, parb, parc = (self.args[arg] for arg in args)
        search_space = [
        [para[idxa],parb[idxb],parc[idxc]]
        for idxa in range(len(para))
        for idxb in range(len(parb))
        for idxc in range(len(parc))]
        return [' '.join([key+' '+str(arg)
               for key, arg in zip(self.args,par_comb)])
               for par_comb in search_space]
```

Listing A.1: Grid Search Hyper Parameters

A.0.4 Tilt and Shift Lense Example



(a) Tilt Shift Optical Diagram



(b) Tilt Shift Qualitative Example

(c) Tilt Shift Qualitative Example Architectural Photography Perspective Correction

Figure A.2: Tilt and Shift Example **Cannon2019**

A.0.5 Aesthetics with Attributes Database (AADB)

This dataset was introduced by **Kong2016** and consists of 10k images. It is available online, however only with Mandarin file details.

S. Kong et al.Kong2016 purpose the datasets as a solution to some the pitfalls of other large-scale datasets, such as class imbalance and the inability to ascertain whether the same user has voted multiple times. AADB was produce by web-scraping from Flickr and random-sampling before abating ground truth from five different individual raters to then compile a confidence score for each attribute.

The MOS scores are treated as a continuous variable rather than being thresholded into a good or bad category.

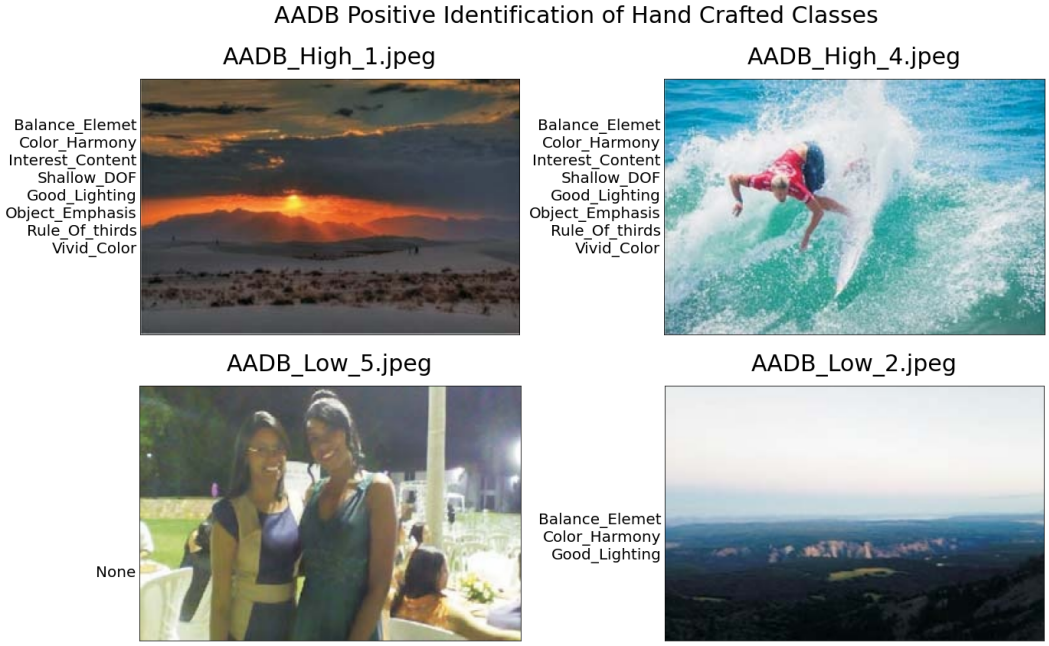


Figure A.3: AADB hand crafted classes

?? shows subjectively that those with a high number of aesthetic attributes are associated with binary classes of high and low aesthetic image quality. Top row images clearly showing high dynamic range, preservation of The dataset is split into high and low and author use. This in combination with aesthetic attributes for algorithmic inference of high or low images based on aesthetic attributes.

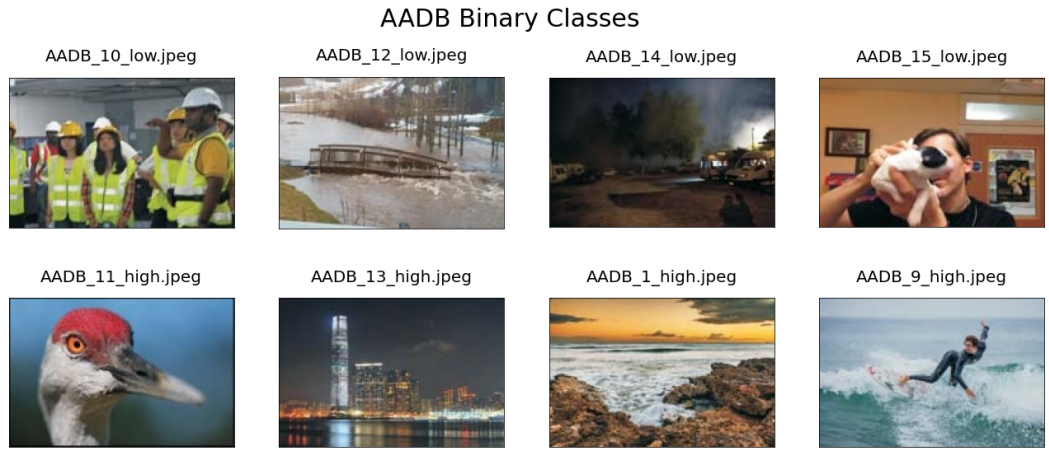


Figure A.4: Examples of AADB data-set binary classes

?? gives some insight into what high and low images 'look' like, and it is clear to the human/subjective eye that even from a small number of images that aspects such as multiple salient objects - where the foreground object is out of focus (top right) - are attributes of low quality images.

However, this also highlights the high level of potential for ambiguity, wherein other images shallow DOF indicated by image features such a bokeh and high levels of blurriness

around are also present in low quality images. Clear visual comparisons appear in figure **??** bottom left and AADB **??** top right, where both images have DOF - however, the bottom left high quality image has a salient object (bird) which is in focus where AADB 15 has a salient object (animal) out of focus.

Ground truth within the dataset was provided by Amazon Mechanical Turk and images within this dataset have controls applied, such as the exclusion of synthetic or heavily edited images and images tagged with qualities *interesting content, object emphasis, good lighting, colour harmony, vivid colour, shallow depth of field, motion blur, rule of thirds, balancing element, repetition, symmetry*. These are showing in **??**, and also provide an insight into high and low quality images.

Aesthetic Rating Online Database AROD

AROD was compiled from images from Flickr between January 2004 and November 2016**Schwarz2018a** and takes note of the number of 'faves' and 'likes' of the images.

They propose a scoring function into high and low categories. This presents a very different distribution of images than is seen elsewhere, and while dataset size is significant, the criteria for scoring data remains somewhat distinct from others where thresholded from a normal distribution.

The ground truth of this dataset is taken from a much larger sample of raters, with an mean of 7k data-points per image. The authors make use of AMT to validate the usefulness of the derived ground truth metric that is computed from uncontrolled user clicks.

The final model is then trained on ResNet50 and produces 75.83 accuracy - it is, however, unclear whether training a binary classifier in this way would perform well on a bench-marking dataset such as AVA, as the authors do not provide evaluation metrics on datasets used elsewhere.

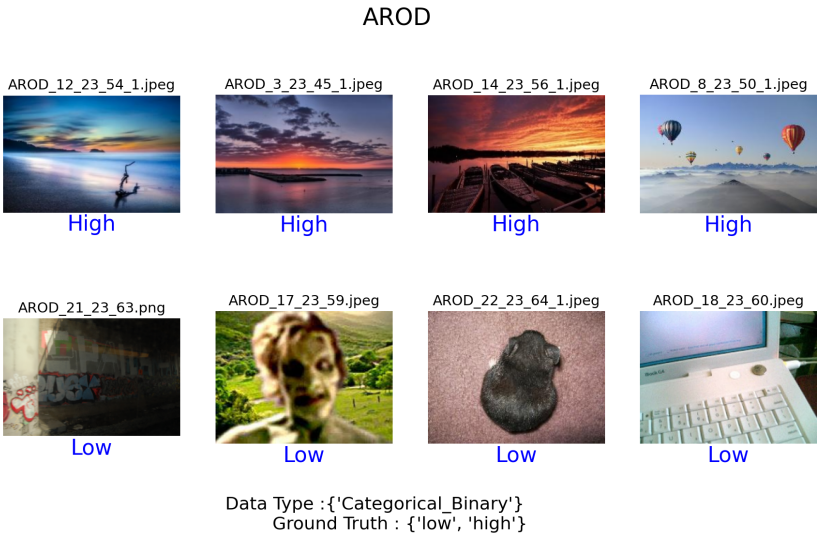


Figure A.5: AROD data-set binary classes top: high, bottom: low

Note centre left, where the salient object is out of focus and the background is in focus in middle bottom left - this is subjectively apparent across datasets where, within a class, there are often images that apparently have otherwise similar visual appearance but where there remains a contradiction of an aesthetic principle (that the salient object should remain in focus).

A.0.6 ALIPR

This dataset represents a unique example of sentiment tagged images. The authors also highlight the pitfalls of using crowd sourced MOS scores for sentiment; this dataset presents a novel approach for inferring individual image aesthetics to augment blunt MOS rated images.

The dataset is compiled by **Datta2008** and is and early example of an IAQA dataset. This was derived using **Li2003** developed by **Wang2002** and is an example of building on existing hand-crafted feature work within IQA.

It was part of hand-crafted feature selection for rating images based on 10 different emotions: *Surprising, Amusing, Pleasing, Exiting, Adorable, Boring, Scary, Irritating, Other, No Feeling*.

The authors aggregate data for descriptive purposes and employ traditional regression analysis rather than for machine learning on their relatively small dataset. The finding, while interesting, also highlights that many of the images either within the category of 'no feeling' were pleasing of boring. This highlights the ambiguity of assessing the sentiment of images.

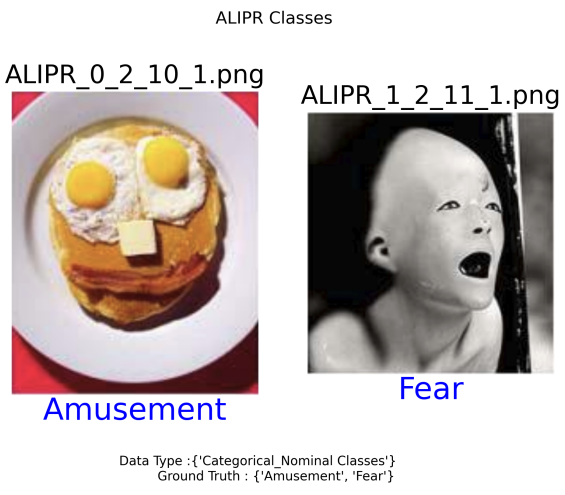


Figure A.6: Examples of ALIPR nominal classes of sentiment tagged images

A.0.7 Aesthetic Visual Analysis AVA

The AVA dataset that is the subject of this thesis is the third largest dataset reviewed within this section, and remains the most well established and canonised example of an IAQA. It is cited within 203 IEEE journals and 305 journals more widely **AVA2012**.

Further, it also maintains the most normal distribution, which is lower than that of other datasets such as PhotoNet. Further still, the distribution of the number of ratings is more approaching normal than the datasets of photo.net, which has large negative skew.

The AVA**Jin2016**, **Hosu2019**, **She_2021_CVPR**, **Lu2015a**, **Redi2015a**, **Chen2017**, **Cui2019**, **Simond2015**, **Kang2020**, **Yang2019**, **Li2020a**, **Jin2020**, **Wu2016**, **Sheng2018**, **Ma2017**, **Liu2017**, **Mavridaki2015**, **Aydin2015**, **Spathis2016** database is the database used within this study, and is to date the largest bench-marked dataset from a single source**Hosu2019**, **She_2021_CVPR**.

The AVA dataset was originally scraped from DP.Challenge.com (the namesake of the earlier database). Each image within the AVA Dataset is ranked from 0-10 with (insert mean number of votes per image).

The images are voted on by challenge participants, and voting can take place both during and after competitions.

A.0.8 Chinese University of Honk - Kong-Photo Quality (CUHK-PQ)

The CUHK-PQ dataset was introduced by**Tang2013a** and consists of 17,673**Tang2013a** ∼ 17,613**Murray2012** high quality images scraped from photo communities and low quality provided by university students with meta data on the score. The dataset also excludes photos with less than 100 votes to ensure a high degree of confidence. Further, the dataset consists of only absolute top and bottom quality classes 1 and 10 where voters on dp.challage.com are able to score images on a scale of 1-10.

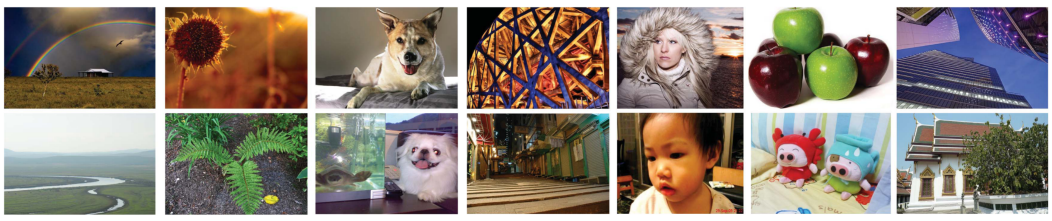


Figure A.7: some examples of high and low quality images form the CUHK-PQ dataset - top is high and bottom is low quality

The CUHK-PQ has a 0.5/0.5 test train split. This dataset was compiled largely with an objective on undertaking machine learning with hand-crafted features.

Others, such as **Lo2013**, **Cui2019**, have used a subset of this dataset to perform IAQA experiments. Images are further sub-grouped into scene categories of:

$$\in \{landscape, plant, animal, night, human, static, architecture\}$$

A.0.9 DP Challenge

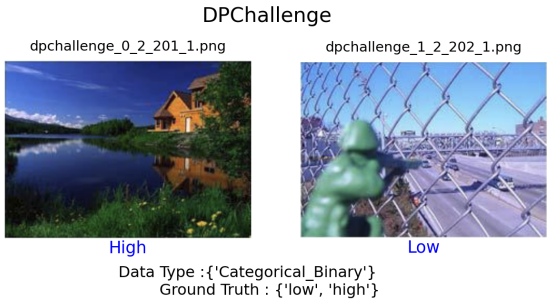


Figure A.8: DPChallange Dataset binary classes **left** high quality **right** low quality

Were amongst the earliest examples of an IAQA dataset**Yang2019** and present the first dataset that is thresholded into binary categories *high* and *low* categories**Datta2008**. This dataset was initially defined in 2008, however has either been used or obtained new images by **Jin2019**, **Dhar2011**, **Wu2011**, **Lou2008**, **Aydin2015**, **Gadde2011**, **Ke2006**, **Gao2015a**, **Nishiyama2011** and others. DP.challenge.com is also the source of the AVA bench-marking dataset **Murray2012**.

To date, there are 330k images on the www.dpchallenge.com and it remains the largest source of images with recorded MOS that are freely available.

Flickr AES

The three datasets that are subsets of available images on the Flickr photo-sharing site are focused on particular fields of study, where researchers have required data on how individuals have related images. There are several so-called Flicker datasets, and the term has been somewhat ambiguously used within the literature.

Yin2012 compile a dataset of 9,600 images with geotagging from Twitter, located from a large auxiliary dataset of 32,00 images. With 8 scene categories, where images are selected from top and bottom categories.

Ren2017 collected Amazon Mechanical Turk (AMT) ratings from 1-5 by five different workers to create ground truth. The 4,737 test images of the dataset proposes the authors are also able to model how individual labels and measure against image datasets.

Chen2016 web scraped 70,00 $\sim$ 90,00 images from 35 identified Flickr groups. They treat each image set separately, and do not provide test training splits or public versions of the dataset. The largest early example of Flickr as a data source is **Cheng2012**, who scraped 80,000 highly rated images for unsupervised learning using hand-crafted features.

Schifanella2015 also compile their own dataset from Flickr to highlight challenges in deriving user produced data from social media sources (such as Flickr) where the number of likes is taken as a proxy for quality.

Image Aesthetic Database(IAD)

To date, this is the largest dataset of IAQA and is derived from DPChallenge (300k) images and w.2 million images form PhotoNet. The mean score image from PhotoNet**Lu2015a** thresholded images from dpchallenge.com and PHOTO.NET.

The threshold boundary for PHOTO.NET is 4.88 and they remove all data within the AVA dataset inter quarterly range (IQR), and are left with only the top and bottom 20% of images. **??** shows examples of high and low quality images after thresholding.

DP.challenge.com has a rating of 1-7 compared to AVA's 1-10, and their whole threshold follows on in principle. Similarly, the distribution of photo.net surfers - for being more atypical to DPChallenge images (this in itself is an interesting and noteworthy phenomenon) and in spite of a much high number of images, show distribution of scores less uniform and with a much higher degree of variability.

while **Lu2014aLu2014a** were able to leverage an increased dataset size for state of the art results at the time of publication, it is noteworthy that many more recent publications have far surpassed their accuracy using only the AVA bench-marking dataset **Zhang2021d**, **Ma2017**, **She_2021_CVPR**, **Kong2016**, **Jin2016**.

Further, the authors make mention of 300k images from DPchallenge; to date, there are 330k images and the date they were rescraped would have therefore included the AVA test images, given that it would not have been possible to obtain 300k images from Dpchallenge in 2016 without including the $\approx$ 20k test images (as there would not have been 90k 'new images' without including them).

Photo Critique Captioning Dataset PCCD

This dataset consists of images scraped from a single source **Chang2017**, **guru** and is based on professionally reviewed of photos where comments are given on general impression, composition, perspective, colour and lighting with a subset of photo, DOF, use of camera exposure and shutter-speed.

Figure **??** shows some particularly high quality images subjectively.

One notable feature of this is the high ratio of comments to image, with $\approx$ 60k captions. Each image is given a numerical rating of 1-10.

Photo.Net

This originally consisted of 3,581**Datta2006** images, however later publications **Joshi2011** augment this or re-scrape images to derive datasets of 20276 net. One distinction is that while many of the datasets are, in practice, thresholded into binary $\in \{good, bad\}$ that Photo.Net has ranked 1-7.

Figure A.9: Examples of 7 AES data-set binary classes

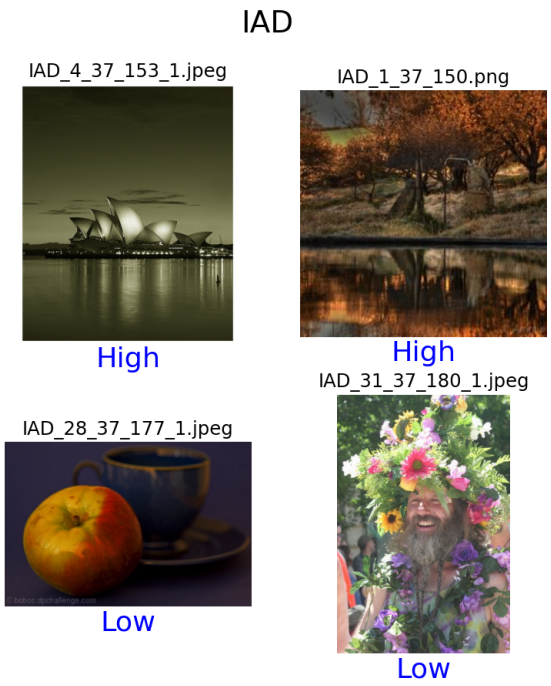


Figure A.10: Examples of IAD data-set binary classes

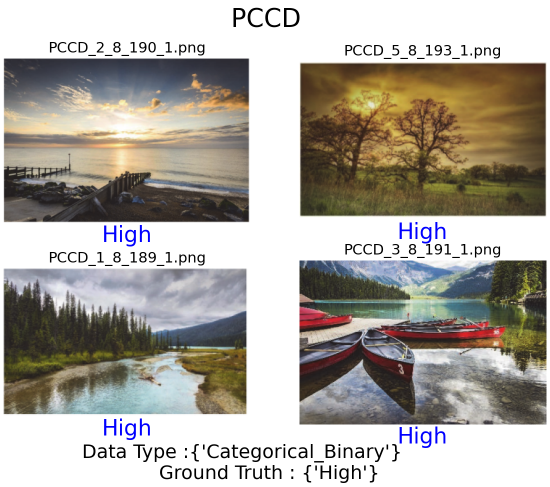


Figure A.11: Examples of PCCD dataset

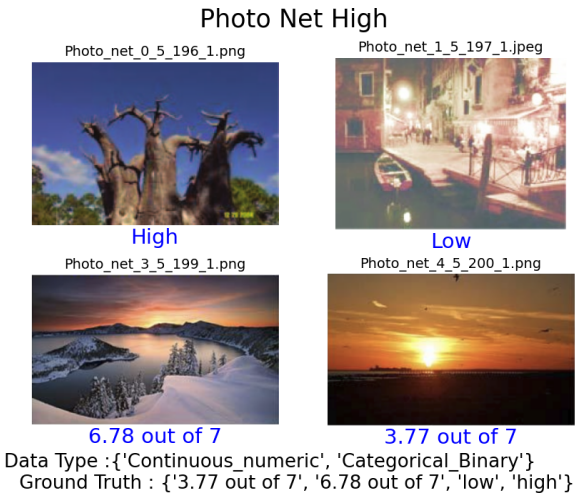


Figure A.12: Examples of PHOTO.NET

Waterloo IAA

This small dataset was compiled for content subjective assessment, while it is not in itself large enough to meaningfully conduct deep learning based approaches.

Liu2017 compile a rich database of aesthetic features that are collated under controlled conditions. The associate publication also demonstrates, for instance, that it is difficult for humans to assess aesthetics from a single factor, but rather that it is in consideration of combinations of factors.

The data was compiled using a computer program where users have a sliding scale, and are able to use a graphics user interface (GUI) in order to capture rating feedback from experimental subjects. The feedback is then used to subset the data for traditional statistical analysis, rather than machine or statistical learning.

This represent a similar approach to AADB datasets in its rigorously paired down subcategories.

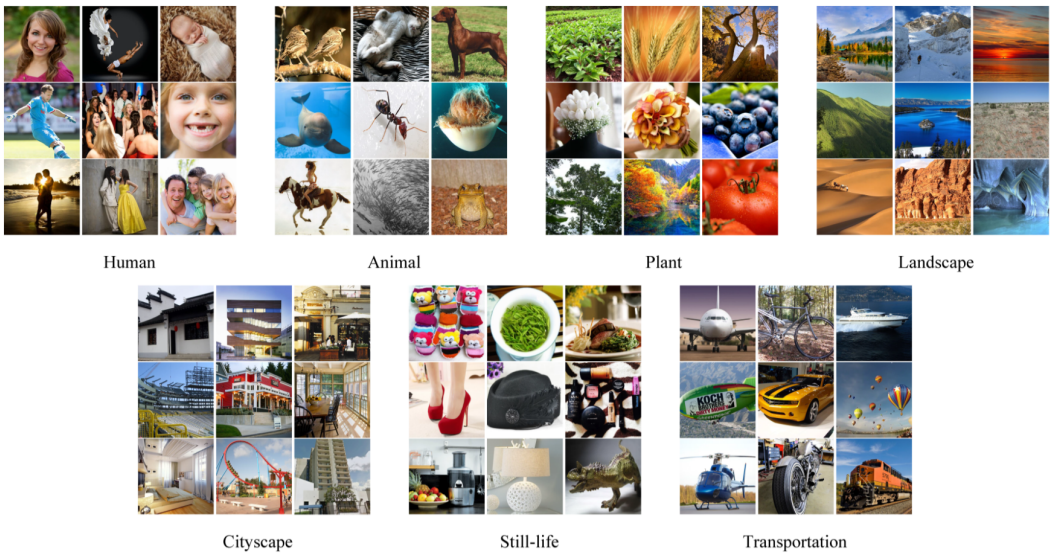


Figure A.13: Waterloo class examples

Liu W.**Liu2017** also apply high levels of scientific rigour to controlling the environment in which the data is captured from subjects, for example calibrating the screen and reporting its results.

A.0.10 Dataset Conclusion

The datasets reviewed here are by no means exhaustive, and the review of IAQA datasets could comprise an entire literature review or survey paper in its own right.

There are fascinating and bespoke examples of user rated databases and retrieval system such as Terra Galleria**Quand-T** which are well canonised**Mayssara2014**, **Hutchison2013**, **Joshi2011** images of a single genre, a single photographer, but with voting data of scores between 1-10.

In addition to deep learning that has been leveraged for evaluation of individual artists such as Andy Lomas**McCormack2021**, both of which are well canonised examples of more bespoke datasets. Some earlier examples of hand-crafted features scrape inames for a single source such as **Cheng2012**.

There are also dasasets on the wider domain of IAQA, such as Chinese Handwriting Aesthetic Evaluation Database CHAED **Sun2015**.

The terminology used, including abbreviations of different databases, are not consistent throughout the literature. This is for three reasons:

1. First, that the source of the databases such as Flickr, DPChallange, or Photo.Net have been used within database nomenclature to indicate datasets, many of which are themselves subsets of other datasets, **Spathis2016** also use the title of a paper as the acronym for a dataset that was created by the authors.
2. Second, and perhaps tangential to the slippage in nomenclature between dataset and source, there there are some inconsistencies in literature for instance **Kanwal2021** use the term Photo web citing **Datta2006** where the term Photo.Net is used.
3. Third, that subsetting of datasets within individual publications, while retaining the super set name, makes the process of effective bench-marking and maintaining the official test train split.

This is challenging, as it leads to potential for inconsistent benchmarking and may serve to dilute the absolute value of results where one subset may be present in test sets of another set with no clear way to ensure that there is not what is in effect a data leak by proxy, with **Lu2015a** as an example of this.

Further, and perhaps in support of evidence of some lack of cohesion within IAQA community, datasets are not always consistently cited for instance **Lo2013** cites **Tang2013a** as a source dataset and then subset this into an new dataset with a different ratio of positive to negative class, which is then cited by **Kanwal2021** as Photo Databse.

This is further compounded by the variability in nomenclature of derivative data sources such as dp.challenge and Flickr. (Further, the dataset size is not always consistently reported for individual dataset.)

One aspect of useful further research would be to provide a consistent high level framework for IAQA dataset and a data schema, which might include a process of using image retrieval to check for duplication's and use of original site image ID's where for instance image ID's are used within image urls .

An aspect of additional complexity is that there exists a high degree of sub-setting or creating new datasets, as can be seen in some of the more recent and smaller datasets. For instance, **Jin2019** upscales the approach of **Chang2017** onto an augmented AVA dataset by web scraping further images for dp.challenge.com.