Utrecht University
Department of Information and Computing Sciences

Glowing Skin, Radiant Beauty

A Deep Neural Network Approach to Detect Facial Glow and Beauty

Diana Gyurjiyan

Supervisors:
Prof. Dr. A.A. Salah
Prof. Dr. M. Smeets
Dr. F. Di Cicco

Second Reader:
Dr. I. Önal Ertuğrul

Rotterdam, 05-2023

Abstract

Facial beauty classification is a relatively new area of research in the field of computer vision, which has been easier to study due to the emergence of social media platforms. The concept of facial glow has not been extensively researched in computer vision, and hence, there is a lack of computational approaches and no available databases or previous studies. This master thesis project aims to address this gap in the literature and real-world application by training an algorithm to automatically classify facial beauty and glow based on the analysis of facial images, using criteria identified from the literature. Additionally, a subset of the CelebA dataset is annotated with facial glow labels in order to create a labelled dataset for the feature *facial glow*.

To limit the risk of Western bias, the influence of cultural and regional differences on perceptions of facial glow are taken into account by using participants for the annotation from different cultural backgrounds. The CelebA dataset is annotated by participants from various countries and regions, providing a diverse perspective on facial beauty and glow.

Additionally, a native language experiment is conducted for the study. The experiment has two important components, namely native language search in the native search engine and usage of a VPN to connect to different local servers across regions. This experiment adds an extra layer of insight into the cross-cultural perceptions of facial beauty and glow.

The primary algorithms for classification in this study are the `VGGNet` models, both VGG16 and VGG19. Multiple algorithms are trained in order to evaluate the performance of deep learning methods. Expanding on these models `Siamese networks` are used to improve the performance and classification. The advantage of the Siamese networks is the ability to compare and classify images based on the similarities and differences, rather than on individual features. This approach enables the algorithm to identify subtle variations in facial features, which can be used as indications of facial beauty and glow. Based on the results the VGG models have significant better performance than the Siamese Networks.

This study contributes to the advancement of computer vision techniques for facial analysis focused, especially on facial beauty and glow, and provides a foundation for further research in this emerging field. Moreover, it also emphasises the use of a more diverse and inclusive approach to the subjective notion of facial beauty and glow.

Acknowledgements

For this research project I would like to express my sincerest appreciation to my thesis coordinator, Prof. Dr. A.A. Salah, for their valuable guidance and support throughout this research process. His profound insights into the subject matter proved to be highly valuable for this project, and his guidance and assistance lead to a successful completion of this research.

Additionally, I also want to express my gratitude to Dr. M. Smeets from Unilever for her contributions to this research project, which have resulted in its improved quality. Her profound expertise and helpful suggestions have been beneficial to this study, and I am grateful of her valuable time and effort.

Finally, I wish to express my appreciation to Dr. F. Di Cicco for her insightful feedback and guidance during the initial phase of the project. Without her multidisciplinary knowledge and expertise, the development of the criteria for both concepts would not have been possible. Her guidance and support has been an integral component of this research project.

"*What beauty is, I know not, though it adheres to many things.*"
— Albrecht Dürer

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

1.1.1 Facial beauty recognition

Facial beauty and glow are important concepts in our society, as they are often associated with attractiveness, health, and overall well-being. Human preferences for facial beauty and beauty in general, develop over time and are influenced by social-cultural context. A few decades ago, beauty standards were previously disseminated to the public using traditional media, like newspapers and magazines. The emergence of social media has increased the reach of beauty standards and their cross-cultural influence. This cross-cultural influence and the exposure of these beauty standards to a broad audience contributed to a more complex and nuanced notion of facial beauty. As a result, the beauty industry is facing challenges in creating effective marketing content. The cross-cultural notion of beauty and glow challenges the one-size-fits-all approach to beauty and glow. Applying a universal approach with the underlying assumption that everyone has the same preferences and values is no longer sustainable. Rather a more diverse and nuanced approach is necessary in order to take into account different cultural perspectives on beauty.

A theory already incorporating the idea of subjectiveness is the Kantian theory. The Kantian theory of aesthetic judgement is a good framework to understand the subjective human evaluations of beauty. According to Kant, *aesthetic judgement* is based on the subjective feeling of pleasure or displeasure that arises from the perception of the form. Applying this framework to the concept of facial beauty and glow leads to the following. Individuals are perceiving beauty and glow based on their own perceptions and experiences. These perceptions are influenced by various factors, including cultural, social and personal norms. Beauty and glow therefore are not merely an objective property of an object or subject, but rather a subjective experience that is influenced by individual interpretations. Hence, it is important to recognise the subjective nature of beauty in order to understand the diversity of perceptions on facial beauty and glow across different cultures and regions. Therefore creating an objective notion can be difficult [Kant, 2000].

The perception of objects in reality is highly influenced by an individuals cultural, social and personal norms. Therefore, judgements of facial beauty and glow are highly subjective and biased. Recent technological advances have enabled the development of algorithms that outperform humans in the classification of facial beauty. These algorithms are widely used for effective and target marketing campaigns. However, the limitation of these algorithms is the bias towards Western beauty ideals. for instance, a search for the term *beautiful face* on Google will results in predominantly images of Caucasian individuals. The key problem is challenging this Western beauty ideal by using cross-cultural influenced dataset.

1.1.2 Facial glow recognition

Previously, there has been a lot of research on facial beauty classification with the use of algorithms. Most of these algorithms are typically trained on large datasets of annotated facial images. The algorithms use different learning techniques, such as learning soft and hard biometrics in order to correctly classify whether a facial image is attractive or non-attractive [Terhörst et al., 2020]. The concept of facial glow is relatively new and maybe more subjective than facial beauty. While there have been studies that investigate the effects of lighting, makeup and skincare on facial glow, there is no clear and objective definition of what constitutes facial glow. Moreover, because of the lack of research in this field, there are still no datasets available annotated specifically for facial glow. The lack of the existence of an annotated dataset for this purpose makes supervised learning impossible.

The main challenge of the classification of facial glow is the broad diversity of interpretations. Some studies define facial glow as the lack of *tiredness* or *aging* [Flament et al., 2020]. A young skin would therefore be a criteria for the classification of facial glow. On the other hand, some scientists suggest a more subjective definition of facial glow, with *genuine happiness* as one of the necessary components [Sheldon et al., 2021]. Facial glow has no agreed-upon definition or standard measure and the interpretation of facial glow can vary per context. As a results, there is no ground truth or reference point for facial glow, which makes classification difficult. However, successfully detecting facial glow leads to improvements in general health care, dermatological diagnoses and effective marketing campaigns for skin-care products. In dermatological diagnoses it can be used as a measure to indicate various skin conditions or diseases.

In this study diverse dermatological and medical papers are used in order to set a ground truth for facial glow. The definition of facial glow encompasses a collection of facial attributes, among which are smoothness, hydration and even skin tone without spots and blemishes.

1.2 Research Approach

1.2.1 Research questions

As previously mentioned the challenge of cross-cultural biases remain in facial beauty and glow detection, despite the technological advances. Thus, the aim of this thesis project is to investigate requirements for the classification of facial beauty and glow while limiting the risk of cultural and social bias. The main research question is as follows:

> How can classification algorithms be developed for facial beauty and glow while incorporating cross-cultural complexity?

The research question can be divided in the following chronological sub-questions:

- **Sub-question 1.** What are the criteria for the classification of facial beauty and facial glow?
- **Sub-question 2.** How is the social-cultural context influencing the requirements of the perceptions of facial beauty and glow?
- **Sub-question 3.** What are existing computational approaches for detecting facial beauty and glow (and how are they influenced by these social-cultural bias)?
- **Sub-question 4.** What are ethical implications of the classification of facial beauty and glow?

The aim of **Sub-question 1** is to identify criteria or requirements for determining facial beauty and glow. A lot of studies have been done on the concept of facial beauty, while facial glow is relatively new without a set ground truth. Therefore, the first step is to create a clear and objective definition, which can be used for the classification. As mentioned previously, the concept of facial glow and beauty are highly subjective and influenced by social-cultural factors.

Sub-question 2 investigates the factors influencing the perception of facial beauty and glow. Social-cultural factors can include historical, geographical, ethical and cultural values. This sub-questions also emphasised the need for a more inclusive and diverse dataset.

Sub-question 3 focuses on the existing machine learning techniques and the applications in facial beauty and glow detection. Different machine learning techniques are compared in order to use the most appropriate models for this study. Namely, the different state-of-the-art models, such as the two `VGG` algorithms. Moreover, not only traditional neural nets are investigated, but also `Siamese networks` in order to improve the performance of the classification.

Finally, it is important to not only focus on the performance of the model but also investigate the ethical implications of the study, which is incorporated in **Sub-question 4**. Automatic classification of facial beauty or glow can lead to certain beauty or glow standards, which should be evaluated.

1.2.2 Research method

This master thesis project is divided into distinctive parts, each aimed to answer one of the sub-questions. The diverse nature of the sub-questions lead to different research methods. The main methods used in this thesis projects are: literature study, modelling and experimental analysis. The literature study is used as a foundation for a clear and objective definition of facial beauty and glow. Additionally, it is used to provide theoretical background and select the right features and techniques for the classification. For a better understanding of the objective definition of facial beauty and glow the literature research reveals underlying objectives of these concepts in both technical terms and everyday explanations. Sub-question 1 is answered by conducting a literature study.

Moreover, the following experiments are executed in this master thesis project:

1. Annotation of the dataset by a diverse and inclusive group of participants.
2. Native language search using local servers.
3. Testing the best performing model on a diverse dataset.

As mentioned before, a facial glow annotated dataset is lacking, which is necessary for the supervised methods using in this master thesis project. Therefore, there is a need for a newly annotated dataset. A subset of the CelebA dataset, including 2000 images, is annotated with the label *Glowing*. This dataset is used to train and test the models and evaluate the performance.

The second experiment is conducting a native language search using the corresponding search engines in different countries across the world. The results of the native language search is creating a new dataset by downloading the image results. This experiment has both a qualitative evaluations and quantitative evaluation.

The last experiment is testing the bias of the best performing model to a more diverse dataset in order to see how the model is performing on facial images of people with a different culture background. The UTKface dataset is used for this purpose due to the high variety of different skin types, colours and forms. The aim of this experiment is to uncover the difference in perception of facial beauty and facial glow for different ethnicities.

1.2.3 Model and Annotation measurements

The performance of the facial beauty and glow recognition pipeline for the automatic classification can be evaluated by several metrics. The following metrics are used in this study by using the classification report:

1. Accuracy: the percentage of correctly classified instances.
2. Precision: the ratio of correct classified positives and wrongly classified examples (false positive).
3. Recall: the ratio of true positives and the wrongly classified examples (false negatives) [Sokolova et al., 2006].
4. F1-score: the harmonic mean of precision and recall [Lipton et al., 2014].
5. Confusion matrix: overview of the wrong and correctly classified labels [Sokolova et al., 2006].

All models are compared to each other using the performance metrics shown above.

For the experiment with the human annotators the Fleiss' Kappa coefficient is used in order to calculate the agreement score between the annotators.

1.3 Thesis outline

This master thesis is structured as follows. In Chapter 2 sub-question 1, 2 and 3 are addressed. The requirements of an objective notion of facial beauty and glow are determined by using a literature study. Additionally, social-cultural influences are also investigated, including their influence on the perception of facial beauty and glow. Finally, different state-of-the-art models are compared, including the `VGG16`, `VGG19` and Siamese networks.

In Chapter 3 the general methodology of the thesis is discussed in more detail. The methodology, results and discussion of the three different experiments are discussed in chapters 6, 4 and 5, for respectively the human annotation, classification task and the native language search. The three experiments are separated in order to increase the readibility of this thesis.

In Chapter 7 the general discussion can be found, followed by the conclusion for all experiments in Chapter 8.

Chapter 2

Literature Study

2.1 Introduction

This chapter includes a comprehensive literature review of the concepts of facial beauty and glow. This chapter also provides insights on how social-cultural factors are influencing, shaping, and changing perceptions of facial beauty and glow, hence contributing to the development of beauty standards. The aim of this chapter is to provide well-defined criteria for both facial beauty and glow in order to enable computational classification possible. These well-defined criteria are necessary for the experiment as explained in Chapter 3.

2.2 The History of Facial Beauty and Glow

2.2.1 History of the Perception of Facial Beauty

Facial beauty has been a popular theme of human interest. Even in ancient Greece mythology, humans were interested in the concept of beauty and created stories about beautiful women causing wars, such as Helen of Troy [Blondell, 2013]. The perception of beauty however does not remain the same, but changes rapidly and is a dynamic concept highly influenced by social and cultural changes. Beauty standards however do not only vary over time periods but also across different regions and different cultures. Despite the dynamic nature of facial beauty, some commonalities can be derived from all the beauty ideals over the years [Konstan, 2014, Thomas and Dixon, 2016].

One of the similarities in the perception of facial beauty is the application of the golden ratio, which is derived from the ancient Greek beauty perception. The ratio is applied to facial features and is defined as the mathematical ratio of 1.618. This is according to some mathematicians and scientists the ideal proportion for aesthetic beauty. The golden ratio can also be found in art, especially during the Renaissance period. Different artists, such as Da Vinci, Boticelli and Michelangelo applied the golden ratio and facial symmetry in their paintings and sculptures. An example of the golden ratio applied in art is the famous painting *Mona Lisa* as can be seen in Figure 2.1 [Livio, 2002, Singh et al., 2019].

Another important requirement of facial beauty throughout history is the smoothness of the skin, which is depicted in Greek and Roman sculptures, but also in drawings of other cultures, such as the ancient Egyptian culture. The smoothness of the skin is also an indicator for facial glow [Romm, 1989, Prokopakis et al., 2013, Zhang et al., 2017a].

There are also large differences in the perception of facial beauty throughout history. Ancient Romans preferred strong defined features, such as a jawline and straight nose, while ancient Egyp-

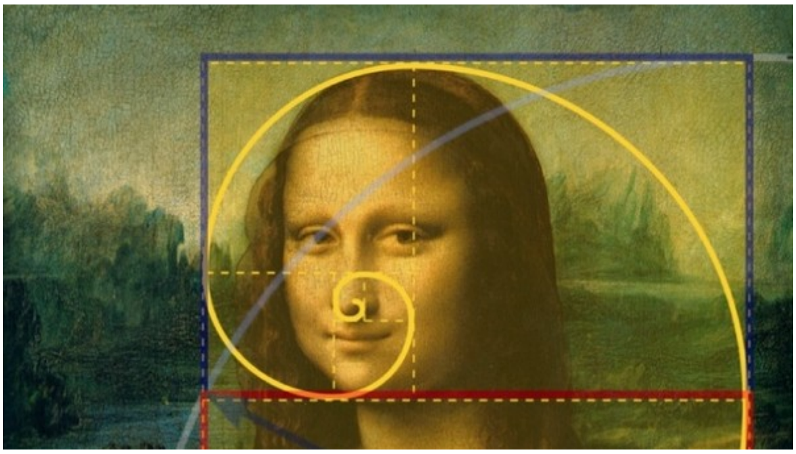

Figure 2.1: The Golden Ratio in Da Vinci's Mona Lisa

tians' perception was focused on a high forehead and arched eyebrows. Over time the perception of facial beauty changed and evolved, especially with the emergence of globalisation. Instead of pale skin, which was an indicator for wealth during the medieval times, a sun-kissed skin became the indicator of beauty and wealth. [Olson, 2009, Wilner, 1931, Heppt and Vent, 2015, Eldridge, 2015, Dimitrov et al., 2023].

From the early 20^{th} century, artists challenged the traditional facial beauty standards when the modernist movement emerged. Artists like Pablo Picasso introduced cubism, changing the perception of facial beauty into more radical and grotesque. In the 21^{st} century, there is more emphasis on natural features and more inclusive perception. Diversity, natural and more inclusive ideas are keywords in the present-day artistic perception of facial beauty [Cohen Jr, 1991, Fernandez, 2001, Samizadeh, 2022].

2.2.2 History of Facial Beauty Practices

Not only the perception of facial beauty has changed and evolved over the years, but also throughout history people have used different beauty practices in order to enhance facial beauty or glow. Different cultures and regions used their own products and ingredients in order fulfil the criteria of facial beauty.

For instance, in ancient Egypt the use of kohl for dark eyebrows and eyelashes was popular, including using different kind of oils and creams in order to moisturise the skin. While in ancient Greece and medieval period, people had a preference for a pale complexion, using white lead powder and chalk to create this. A pale complexion was a sign of wealth, due to the fact that people with a tanned skin, often worked in the field under the sun. In China a similar perception of facial beauty was considered, using crushed pearls and jade to create a pale glowing skin [Shafik and Elseesy, 2003, Westmore, 2001, Yu et al., 2017, Eldridge, 2015].

In the last few years, the beauty industry has become highly dependent on social media, which includes more targeted advertisements. Additionally, society does not accept the one-size-fits-all approach, due to the complexity of nowadays societies (including a mixture of ethnicities). Instead of using chemical products, women are more nudged into using chemical-free creams, lotions, and serums in order to moisturize and protect the skin. Makeup is more focused on enhancing and strengthening natural features instead of creating a mask.

2.2.3 Facial Beauty and the impact of Technology

Another important change in the history of the perception of facial beauty is the impact of technology. The impact of technology can be distinguished in two parts: medical and photography.

Before, the 19^{th} century the only way of portraying and capturing beauty was by using paintings or sculptures. Nowadays, photography enables people to capture facial features in a different way. With the rise of social media, the rapid increase of images also led to a more diverse and inclusive beauty perception.

Before social media, images were shown in magazines and billboards. Hence, the images were displayed only from the creator's intention with limited input from society [Eggerstedt et al., 2020]. However, now social media is taking over, which means that the idealised versions of beauty are unattainable due to the high inference of society and preferences [Henriques and Patnaik, 2020]. The one-size-fits-all approach is not sustainable any more in such a diverse society. Traditional Western beauty standards are therefore not the only criteria in the perception of facial beauty or glow [Laughter et al., 2023, Dimitrov et al., 2023, Oranges et al., 2016, Larrabee, 1997].

Technology also increased worldwide travel, which lead to globalisation. Due to the globalisation cross-cultural interaction is possible and has a great impact on the perception of beauty. The perception of facial beauty is influenced by other cultures and a new perception is born. An example is the growing popularity of Korean beauty ideals and standards. The benefit of globalisation is the rise of a more diverse and inclusive perception of beauty. For instance, a few years ago not all skin shades for foundation were available, while nowadays a range of different colours are available for public (see Figure 2.2) [Palma et al., 2017, Koon, 2021, Samizadeh, 2022].

Brands become more and more aware of the need to represent every skin tone, body type, and beauty ideals [Navarro, 2020]. These changes are not only seen in product offerings but also in marketing campaigns. One notable marketing campaign is for example hiring models with *Vitiligo*, a rare skin disease that leaves the patient with uneven skin colour [Sisti et al., 2021].

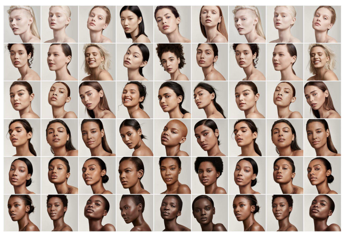

Figure 2.2: Diversity and more Inclusive Fenty Foundation

2.3 Objective definition of Facial Beauty

The perception of beauty has commonalities throughout history and different cultures. Based on literature and previous studies the following objective criteria are chosen to be discussed in this chapter:

1. Golden Ratio: the mathematical ratio between two features.
2. Facial Symmetry: the symmetry level of the face.
3. Skin Quality: absence of acne and other irregularities of the skin.
4. Facial Features: the form of facial features.
5. Facial Averageness: the averageness level of the face.

2.3.1 Golden Ratio in Facial Beauty

One of the most common prerequisites of facial beauty is the golden ratio, which is widely used in various fields such as art, cosmetic surgery, and architecture. The golden ratio is also known as the divine proportion and refers the ideal proportion between two features. According to some the ratio creates a sense of harmony and balance, which is aesthetically pleasing to the human eye, and can be found in both living and non-living things. In the context of facial beauty it can be applied to all facial features, including the width of the nose, the height of the forehead, and the distance between the eyes [Yalta et al., 2016, Dunlap, 1997, Posamentier and Lehmann, 2011, Gunes and Piccardi, 2006a].

The ratio, 1.618 is called the golden mean or golden ratio [Dunlap, 1997]. Applying this ratio to the width of the face results in the height being 1.618 times the width in order to satisfy the golden ratio as can be seen in Figure 2.3 [Hassaballah et al., 2013].

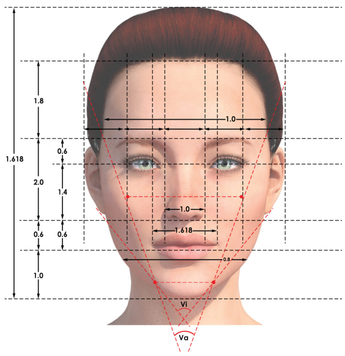

Figure 2.3: Golden ratio for Facial Features

The golden ratio however has its own limitations and therefore cannot always be used as a requirement for facial beauty. First of all, facial beauty differ across regions and cultures, due to differences in appearances across different cultures. The golden ratio is based on mathematical proportions and does not take into account other features which may influence the perception of beauty, such as cultural or personal influences. Some scientists have discovered that the golden ratio is not applicable to Malaysian faces, due to the different form of the face [Alam et al., 2015]. However, the non-applicability does not mean that these faces are not beautiful, but it simply means that the golden ratio is not an universal standard and cannot be applied to all faces. The golden ratio has an internal bias, which can raise risks in facial beauty classification. Relying solely on the golden ratio results into a narrow and rigid definition of beauty and unique features are ignored, which can make a person's face beautiful in its own unique way [Alam et al., 2015, Packiriswamy et al., 2012].

Additionally, the golden ratio is mostly applied to a two-dimensional image. In reality faces are not two-dimensional, but three-dimensional, which means that using this method alone can lead to inaccuracies. The golden ratio may not apply in the case of a three-dimensional face due to different calculations [Zwahlen et al., 2022, Jang et al., 2017].

The golden ratio can nevertheless be a very useful reference point for training neural networks to classify facial beauty. Many studies have used the golden ratio as an input to classify the facial beauty. For example, according to the study of Iyer et al. human judgements were consistent and correlated with the facial ratios. The researchers however remarked that there were other elements of variations, which existed as secondary features, such as sexual features and facial health. Combining these features resulted in a better performance of the model [J Iyer et al., 2021].

2.3.2 Facial Symmetry

Another important indicator for facial beauty is facial symmetry. A face is symmetrical when the right and left side of the face are nearly identical. The animal brain, which includes the human brain, is primarily attracted to symmetrical faces of their mates [Perrett et al., 1999].

When the facial symmetry is too high, which can be the case in deep fake images, the point of the *uncanny valley* will be reached (see Figure 2.4) [Thornhill and Gangestad, 1999]. This effect occurs when the images are too symmetrical, making these images look almost human. Because the images are nearly human, but not human or non-human enough, it leaves a feeling of unease or even fear rather than attraction [Zhang et al., 2020, MacDorman et al., 2009, Perrett et al., 1999].

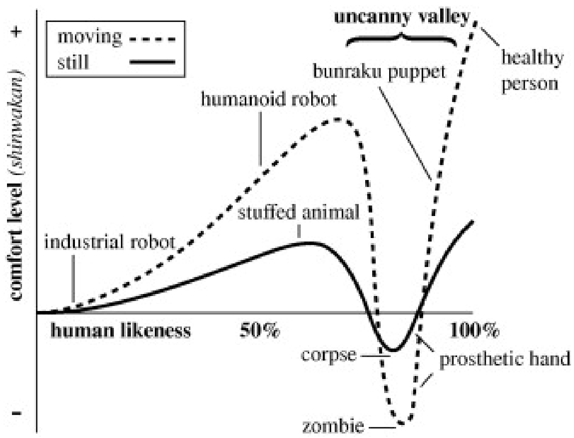

Figure 2.4: Graph of human-like appearances and comfort (Figure from [MacDorman et al., 2009])

Other studies have suggested to not focus on the symmetry of the face, due to the problems described above, but rather focus on the asymmetry of the face. Facial asymmetry, symmetry and average symmetry are shown in Figure 2.5. The asymmetry then can be calculated by the sum of all difference between the feature point. Studies have shown that less asymmetrical faces, are more considered more attractive, healthy, friendly and sociable. Facial symmetry therefore is not only an indicator for facial beauty, but also for various other aspects of the individual such as health and personality [Fink et al., 2006, Verhoeven et al., 2016, Dobai et al., 2018, Huang et al., 2013, Rhodes et al., 2001b, Harguess and Aggarwal, 2011]. Facial symmetry is most often associated to other positive traits, such as intelligence, trustworthiness, likability and better health [Rhodes et al., 1998, Rhodes, 2006, Little et al., 2011, Grammer and Thornhill, 1994].

One study has studied the relationship between facial symmetry and attractiveness in a sample of 147 female images. The study found that women with higher facial symmetry were rated as more attractive by both male and female participants. Another study found a similar results when analysing the relationship between these two aspects between female and male images. Facial symmetry was highly correlated with attractiveness regardless of the gender [Rhodes et al., 1998, Wei et al., 2022, Noor and Evans, 2003, Eisenthal et al., 2006, Liu et al., 2016].

Figure 2.5: An Example of Facial Symmetry and Facial Asymmetry (Figure from [Lee et al., 2021, Fink and Neave, 2005])

2.3.3 Skin Quality

Another indicator of facial beauty is the quality of the skin, due to the fact that a good skin condition is often associated with a good health.

Many studies have proved that skin condition is important for the perception of facial beauty. One of the reasons is that good skin condition is often perceived as more fertile for both female as male faces. Flawless skin without acne or the absence of facial hair is associated with female facial beauty. These are called sexually dimorphic features, which is associated with female and male fertility [Fink et al., 2001, Fink and Neave, 2005, Morris, 1967, Van den Berghe and Frost, 1986].

The correlation between the perception of skin condition and health is not a coincidence, due to the fact that the skin can be affected by unhealthy behaviour, such as lot of sun exposure, smoking and wrong diets. Wrinkles and roughness of the skin can be caused by long-term exposure of the sun. A diet which is high in sugars and fats can cause acne on the skin. Hence, someone's health can indeed be derived from their skin condition [Flament et al., 2013, Genovese et al., 2017, Naidoo and Birch-Machin, 2017, Humphrey et al., 2021].

2.3.4 Facial Features

In combination with the other indicators for facial beauty, facial features itself are also important for the perception of this notion. *Attractive* facial features depend on the ratio and shape of the features itself. People with certain proportions are generally perceived as more attractive.

For instance a smaller face with a larger forehead are considered attractive by females, while other features include having a pronounced chin and fuller lips. The preference for the right facial proportions or features differ per culture. Asian perception of female facial beauty includes having a more prominent jawline, while Western female perception include fuller lips. Even within regions itself the beauty perception regarding facial features can change amongst sub-regions [Naini et al., 2006, Fan et al., 2012, Kashmar et al., 2019, Liu et al., 2019, Cunningham, 1986, Kashmar et al., 2019, Broer et al., 2012, Yip et al., 2019, Menon, 2019].

Additionally, the most important requirement for the facial features is having the right proportion. A nose that is proportional in size compared to the face is considered more attractive than smaller or larger noses. The same applies for the eye size which also should be in balance with the rest of the face. Too big or too small eyes are not considered attractive. The width of the face is an important feature to consider in combination with the width of the nose or eyes [Naini et al., 2006, Fan et al., 2012, Pallett et al., 2010].

2.3.5 Facial Averageness

The last indicator for the perception of facial beauty is averageness. Facial averageness is a well-studied indicator in the field of facial beauty. It refers to the degree to which the face resembles average features of the population. This beauty ideal is highly depended on culture or region and therefore is a good and reliable indicator for facial beauty. Many studies agree on the averageness effect, which means that a more average face is considered more attractive by the participants.

Averageness of the face is a common indicator for facial beauty across different cultures. A study done on Asian, Chinese and Japanese, and Caucasian faces resulted in the conclusion that averageness is highly correlated with the perception of attractiveness. Increasing the averageness of individual own-race faces resulted in an increase of the attractiveness score. When the averageness was decreased the images were considered less attractive[Rhodes et al., 2001a].

Various experiments have explored the effect of facial averageness on the perception of facial beauty. The preference of facial averageness may have evolved due to genetic heterozygosity. Genetic heterozygosity is an indicator that an individual has inherited different versions of the parental genes. It is often associated with a higher resistance to parasites, a better immune function and a lower risk of genetic abnormalities. The preference for facial averageness is therefore not only a social or cultural influence, but can also have an evolutionary base. Facial averageness implies a genetic diversity and potential health benefits for the offspring [Thornhill and Gangestad, 1993, Halberstadt and Rhodes, 2000, Johnston and Franklin, 1993, Perrett et al., 1994].

Average faces are generally more attractive, but some non-average features such as fuller lips, narrower jaw and larger eyes can increase the attractiveness of the face. An example of an average face with non-average features is shown in Figure 2.6 [Alley and Cunningham, 1991, Perrett et al., 1994, Fink and Neave, 2005].

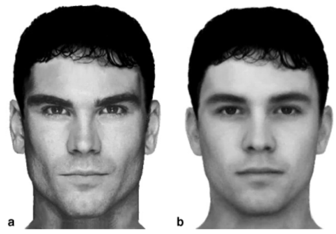

Figure 2.6: Facial images with applied averageness (Figure from [Fink and Neave, 2005])

A recent study examining the asymmetry of attractiveness perception has offered an important relationship between facial attractiveness and the hemispheric specialization [Zaidel et al., 1995]. The analysis of Kagian et al. has revealed that symmetry is strongly related to the attractiveness of averaged faces, but is not the only factor in the equation of facial beauty. The half of the image-features relate to the rating of averaged composited in a similar manner as symmetry. The preference for averages faces was found even when the effect of symmetry was neutralised [Kagian et al., 2008].

Rubenstein et al. discuss a morphing technique to create mathematically averaged faces from multiple face images. In their experiment human subjects showed a preference for averaged face composites even when the effect of symmetry is controlled for [Rubenstein et al., 2002]. Similarly, in the experiment of Kagian et al., the effect of symmetry was neutralised by using only perfectly symmetric component faces which yielded perfectly symmetric composites [Kagian et al., 2008].

According to another study from Kagian et al. a common taste for facial attractiveness are a part of our primary constitutions [Kagian et al., 2008]. According to some psychologists, even infants like to look at a more attractive face [Langlois et al., 1987]. Grammer et al. investigated symmetry and averageness of faces, and concluded contradictory to Perrett et al. that symmetry was more important than averageness in facial attractiveness [Grammer and Thornhill, 1994]. Hence, the use of either facial symmetry or averageness should be combined in order to find the best classification for facial beauty.

Attractive faces might be pleasant to look at since they are closer to the cognitive representation of the face category in the mind. Halberstadt et al. have further demonstrated that not just average faces are attractive but also birds, fish, and automobiles become more attractive after being averaged with computer manipulation [Halberstadt and Rhodes, 2003]. A third view suggests that facial attractiveness originates in a social mechanism, where preferences may be dependent on the learning history of the individual and even on his social goals [Zebrowitz and Rhodes, 2002].

2.4 History of Facial Glow

The history of facial glow is not as extensive as the history of facial beauty. Facial glow is often used to describe facial radiance or luminosity of the skin. This is often associated with good health and vitality. Facial glow is often related to facial beauty and therefore not much literature can be found for this topic. The history of facial glow in this section is therefore a summary of the limited existing literature.

2.4.1 History of the Perception of Facial Glow

In ancient civilisation attractiveness of the face was not only determined by a set of facial features, but also a glowing complexion. This glowing complexion was associated with wealth and status, due to a lack of exposure to the sun and outdoor labour. In other cultures, such as India, facial glow was an indicator for internal health. The practice of Ayurvedic was introduced, which emphasises the importance of a good internal balance. Other Asian countries hold a similar belief in the perception of facial glow [Winkler, 1990, Scott, 2016, El-Kilany and Raoof, 2017, Loughran, 1991, Yadav and Yadav, 2015, Vats and Maurya, 2022].

In present day, a similar approach as the Ayurvedic practice is used. The ideal of facial glow is associated with a healthy lifestyle and skin condition. Skincare products and treatments hydrating the skin and promoting radiance are highly popular. Moreover, The old Indian methods are reintroduced, and many people believe that glowing skin can be achieved by eating the right food.

Another approach to create facial glow is the transition of traditional foundation to nurturing foundations. Withing the make up industry worldwide, traditional foundations are replaced with foundations containing skincare ingredients and therefore giving a more glowing look to the face. The new line of foundation can not only be used for creating an even skin, but also hydrating, exfoliating and protecting the skin. Big beauty brands, such as Dior and Chanel, are already introducing this new foundations [Erdman and Poutahidis, 2014, Zeppieri, 2022, Padgett, 2015, Miller, 2022, Andini et al., 2022, Khanna and Datta Gupta, 2002, Kim, 2021].

2.4.2 Scientific Revolution and Facial Glow

The scientific revolution is major factor influencing the understanding and improvement of facial glow perception. Major scientific techniques such as dermatological and medical research are providing a deeper understanding of the physical aspect of facial glow. Scientific research in combination with new technology have allowed us to gain a deeper understanding of biological and environmental factors that affect the skin's appearance and health. These factors include:

1. Diet
2. Hydration
3. Sleep
4. Stress Management [Pirello, 2001, Kryger, 2023, Ramanigopal, 2015]

Technological advances are used in order to develop products with the right ingredients to stimulate or avoid these factors, but also to nourish and protect the skin in order to promote a healthy glow.

Another important benefit of scientific research gaining understanding not only of the beneficial, but also harmful effects of the environment. One of the examples is popularity of tanned skin, in order to get a natural glow. It is common knowledge that UV radiation and pollution have a negative effect on the appearance and health of the skin. This has led to an increase in protective skincare products, such as foundation with SPF to protect the skin from harmful UV radiation [Sveikata et al., 2011, Abolhasani et al., 2021].

In addition to advances in skincare research and technology, the scientific revolution has also had an impact on the beauty industry more broadly, with a growing emphasis on evidence-based skincare and a shift away from misleading or unproven claims. This has led to a greater understanding and appreciation of the importance of healthy, well-nourished skin for achieving a natural, radiant glow.

2.4.3 Media Evolution on Facial Glow

Finally, the advances of social media and increased use of filters and editing tools have made an significant impact on facial glow. Platforms such as TikTok and Instagram have popularised the use of filters and effects that enhance facial features and create a glowing effect on the skin [Zheng et al., 2021].

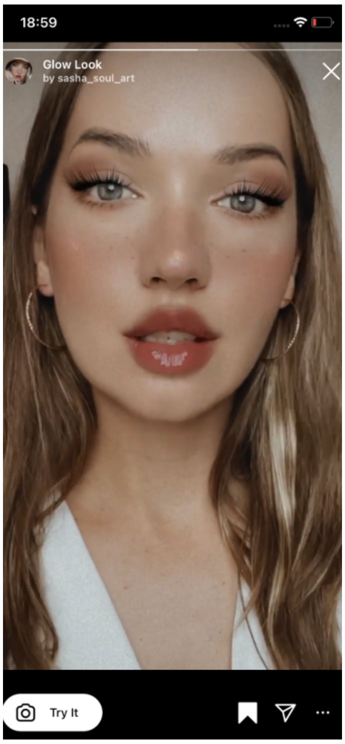

Figure 2.7: Facial Glow filters on TikTok

One popular example is the *Glow* filter on Instagram, which smooths the skin and adds a luminous quality to the skin complexion. In Figure 2.7 the Instagram Glow filter is shown. Based on this image, we can conclude that not only luminosity is added, but also small make-up touch up, which means that in social media discourse facial glow is associated with make-up. Similarly, other social media platforms, such as TikTok and Snapchat have their own *Glowing skin* filter. These filters and effects have become increasingly popular among young users. Companies, such as Clinique, uses these insights in order to better target and create marketing campaigns. Most of the young people using this filters want to enhance their appearance and create a more polished and flawless image [Susilo et al., 2022, Youn, 2019].

These filters do not only have positive sides, but are also criticised for promoting unrealistic beauty standards and contributing to self-doubt in young female users. Some experts have raised concerns about the potential psychological impact of constantly seeing heavily edited images, which are highly realistic (given the high quality of the images) [Harrison and Hefner, 2014, MacCallum and Widdows, 2018, Tiggemann and Slater, 2017]. Users are pressured to adjust themselves to the unrealistic beauty standards, while these cannot be achieved in reality. An example of the unrealistic image is shown in Figure 2.8. In Figure 2.8 a celebrity, who is known to use editing tools and filters, is shown before and after applying the filter.

Figure 2.8: Instagram Images With and Without Filter.

2.5 Objective definition of Facial Glow

While facial glow being a very subjective notion throughout history some commonalities can be derived. In this section objective indicators are discussed in order to make computational classification possible. The following requirements for facial glow are discussed in this chapter:

1. Luminosity of the skin: the overall radiance of the skin, especially in the T-zone.
2. Smoothness of the skin: absence of wrinkles and imperfections of the skin.
3. Hydration of the skin: the hydrated appearance of the skin.
4. Evenness in skin-colour: evenness in colour without any hyperpigmentation of the skin.
5. Elasticity of the skin: ability of the skin to stretch and return.
6. Visibility of the pores: the absence of blackheads and facial pores.

2.5.1 Luminosity of the skin

The first indicator of facial glow is the skin luminosity, which is a measure of the overall radiance of the skin. This is often seen as a key indicator of a healthy skin and is influenced by different factors. The factors include:

1. Skin Hydration
2. Blood Flow
3. Presence of certain Molecules (Carotenoids and Melanin)

There are different methods to measure the skin luminosity, such as reflectance spectrophotometry and perception studies. Reflectance spectrophotometry involves shining a light on the skin and measuring the amount of reflected light. This method is more difficult to execute due to the usage of specialised instruments, which can measure different wavelengths of light. The comparison of these different wavelengths results in an estimation of the amount of light absorbed by different skin components, such as melanin and haemoglobin. These molecules can be indicators for the overall luminosity of the skin [Groathouse et al., 2010, Shah and Chew, 2018, Mohamad et al., 2014].

Another method to measure luminosity of the skin is perception studies using subjective rating scales with human participants. There have not been many (perception) studies on the concept of facial glow. Subjective rating scales, in comparison to reflectance spectrophotometry, involves having participants rate the perceived luminosity of the skin. These subjective annotation are more prone to bias and variability compared to more objective measures. However, they can provide valuable information about the perception of facial glow instead of the objective measurements of the physical appearance of the skin.

Flament et al. is an example of a perception study, in which participants were shown images of Chinese, Japanese and South African women. This study proved that skin texture, ptosis (sagging) and pigmentation signs were the most common facial signs associated with a decrease in facial glow. Additionally, the decrease in facial glow was correlated with increase in the perceived age. For South African women pigmentation signs and cheek skin pores were the most prevalent signs in the perception of facial glow and age. Additionally, the Chinese and Japanese annotators found pigmentation signs as an important indicator of the decrease in glow [Flament et al., 2021].

Another association with facial glow is attractiveness. People with glowing faces are often perceived as more youthful, healthy and therefore more attractive. According to Ikeda et al. facial images with radiance on the entire face were rated as the most youngest appearing faces. The study also found out that the radiance of the face influences the attractiveness rating, especially when the radiance on the T-zone (as depicted in Figure 2.9) is higher compared to other areas [Ikeda et al., 2021].

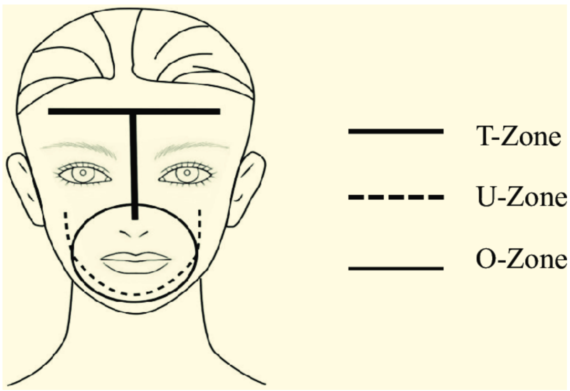

Figure 2.9: Dermatological Facial zones (Figure from [Wahab et al., 2022])

2.5.2 Smoothness of the skin

The smoothness refers to the evenness of the skin texture. A smooth skin is generally associated with youthfulness and a good health. A method to measure the smoothness of the skin is to measure the roughness of the skin.

Rough skin if often associated with signs of ageing (such as wrinkles), sun damage or skin conditions like acne, which can distract from the overall appearance. In contrast, a smooth skin on the other hand is mostly associated with a healthy skin with no acne [Watson et al., 2016]. The association with healthiness, youthfulness and vitality are often found attractive by many people. Additionally, smooth skin is often more reflective and therefore appears to have more luminosity, which can further enhance its beauty. [Sakano et al., 2021, Little et al., 2011].

A rough skin is mostly related to an unhealthy appearance of the skin. An example of a rough skin with a unhealthy appearance is shown on Figure 2.10. The unhealthy appearance of the skin can be caused by numerous factors, among which are:

1. Imperfections: when the skin is free from acne or other facial imperfections, such as scars.
2. Wrinkles: are mostly linked to less good skin texture and are mostly linked to age or lifestyle [Flament et al., 2021].

3. Oiliness: a moderate amount of oiliness on the skin is necessary to have a more moisturised, supple and youthful skin. However, when the oiliness level is high the skin is perceived as greasy and therefore unhealthy [Guinot et al., 2001].

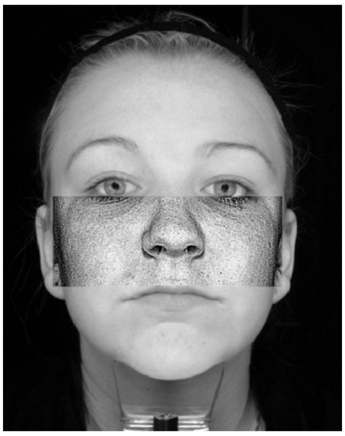

Figure 2.10: Facial images with rough skin condition (Figure from [Fink and Neave, 2005])

2.5.3 Hydration of the skin

Hydration is considered to be a crucial factor in maintaining skin health and beauty. As the skin is the largest organ in the body and serves as a barrier against external harmful influences, the skin is constantly exposed to pollution, UV-lights, and many other damaging influences. Therefore, it is important to maintain an adequate hydration level in order to prevent skin damage and preserve the natural radiance [Kim, 2021, Matsubara, 2012].

The presence of water in the skin cells is crucial for maintaining a optimal skin health and function. Adequate hydration levels ensure the nourishment of the skin and preserving the protective barriers. Moreover, hydrating the skin can also decrease the presence of fine lines, wrinkles and other signs of ageing.

2.5.4 Evenness in skin-colour

An even skin colour is like the other indicators of facial glow associated with health. The evenness of the skin colour can be influenced by many factors, including genetics, sun exposure and skin damage. Uneven skin colour can be caused by skin deceases, such as vitiligo, but also due to extensive sun damage, ageing or inflammation, causing hyperpigmentation, redness and skin discolourations.

Moreover, an even skin colour indicates also a proper circulation of oxygenation of the skin, causing a more youthful and radiant appearance. The healthy blood circulation carries important nutrients and oxygen to the skin, which promotes collagen production, cell turnover and overall skin health [Du et al., 2022, Matsubara, 2012, Ikeda et al., 2021, Haeri et al., 2019].

Colour variation is often associated with ageing, stress and medical conditions, such as rosacea or eczema. The evenness of the skin colour is a result of a healthy lifestyle, including proper sleep and a balanced diet [Shah et al., 2022].

Evenness in skin colour can also indicate the visibility of pores. Pores are a common skin condition enlarged or clogged pores make the skin appear rough and dull, diminishing the overall radiance of the face. Pores are small openings in the surface of the skin, allowing oil, sweat and many other substances to pass through. When pores get filled with dirt, oil and dead skin, they become more visible and make the skin appear rough and uneven. Clean pores that are free of debris are less noticeable, making the skin appear more smoother and even in skin-colour [Flament et al., 2013, Flament et al., 2021, Ikeda et al., 2021].

2.5.5 Elasticity of the skin

Elasticity is the ability of the skin to stretch and return to its original shape. Younger skin is more elastic than older skin. The latter is less supple and prone to wrinkles and sagging. Therefore, elasticity, indicated by the skin age, can be an indicator for facial glow.

The medical causes of skin elasticity are linked to the production of collagen and elastin. These two proteins play a crucial role in preserving the firmness and suppleness of the skin. Over time, the natural production of the two proteins decrease, causing a loss of elasticity in the skin and increasing the amount of wrinkles and fine lines. The presence of skin elasticity indicates a younger and more even skin, which contributes to the perception of radiant appearance [Mukherjee et al., 2011, Maske et al., 2019, Goldie et al., 2021].

2.5.6 The limitations of the concept facial glow

Facial glow, such as facial beauty, is also a subjective quality. One person can perceive a face being *glowing*, while another person would not agree. the ambiguity of the perception of facial glow complicates the objective measurement, which makes the automatic classification with the use of machine learning models complicated.

Moreover, facial glow is influenced by various factors, such as cultures, practices but also technological advances. The factors influencing facial glow include skin, tone, texture of the skin, but also the elasticity. These factors interact with each other in complex ways. As with facial beauty, facial glow can also change over time due to external and internal influences. Factors influencing the facial glow perception is for instance tiredness. Facial glow is therefore not a permanent attribute of the face, such as the shape of the nose or eyes are. Human faces and the attributes of facial glow changes over time.

All these limitations together lead to a subjective notion of facial glow, which is not easily objectified. Objective criteria however are important for classification tasks. The difficulty of creating objective criteria complicate the classification for machine learning models [Sakano et al., 2021].

2.6 Computational approach of facial beauty

2.6.1 Pattern recognition and Facial Beauty

Facial beauty prediction is becoming more popular in science due to the use of pattern recognition and new machine-learning techniques. Scientists try to achieve the automatic classification of facial attractiveness by using computational models. The use of computational models allows scientists to extract meaningful features from facial images and create predictive models. The models learn from large datasets and predict the perceived level of attractiveness with a good performance [Gunes, 2011, Bougourzi et al., 2022, Cao et al., 2020, Zhai et al., 2020, Saeed and Abdulazeez, 2021].

When automatic classification of facial attractiveness is established, it allows the development of automated systems to predict the perceived facial attractiveness level based on the facial features [Saeed and Abdulazeez, 2021]. This automatic classification can be useful in various applications, such as makeup and fashion industry and the development of personalised cosmetic products and procedures [Zhang et al., 2016b].

2.6.2 The Importance of Normalisation

Another method applied to the automatic classification of facial beauty is the normalisation of the facial images. The normalisation of images is done in many studies. Normalisation of the images is an important procedure in facial beauty detection. Using normalisation helps to remove the unwanted variations in facial images that interfere with the accuracy or the learning of the beauty classification algorithm.

Facial images can have variations in illuminations, pose and facial expressions. These variations can influence the performance of the model and lead to incorrect or inconsistent beauty ratings. Normalising the images reduces the variations or eliminates the variations. Normalisation results in a more standardised set of images which lead to better model performance[Eisenthal et al., 2006, Ekmekji, 2016, Xu et al., 2017, Perrett et al., 1999, Saeed and Abdulazeez, 2021].

Normalisation techniques can vary, but include resizing, cropping and adjusting the contrast and brightness of the image. One example is the study of Perret et al. The researchers normalised the images with respect to the facial length in order to limit the background bias and the facial space of the pictures [Perrett et al., 1999, Grammer and Thornhill, 1994, Kagian et al., 2008].

According to the study of Salah et al. another method is to use a perceptual colour space, which can improve the accuracy in age estimation, therefore it can be fruitful to use this perceptual colour space method for the improvement of the classification of facial beauty too. Perceptual colour spaces reflect on how humans perceive colour, including colour consistency or uniformity. By using a perceptual colour space the colour information in facial images are represented more accurately. This can improve the performance of the model [Dibeklioğlu et al., 2014].

2.6.3 Feature studies on facial beauty

There have been many studies on the feature extraction of facial beauty. One way to extract the features is the use of the geometric features. Kagian et al. used this method in order to calculate the smoothness of the skin. The smoothness of the skin was calculated using a edge-detecting algorithm combining the distance-vector and the slope-vector. This resulted in a vector representation of 6972 geometric features. Another study, Zhang et al. used also the geometric feature subtraction method to calculate the geometric distance between the feature points and ratio vectors. The vectors were a combination of low- and high-level features and performed a feature selection. High-level features were age, sex and makeup, while low-level features included landmarks, shapes and structures [Kagian et al., 2008, Zhang et al., 2011, Zhang et al., 2016a].

Machine learning methods can predict attractiveness ratings by learning the mapping from the facial images to the attractiveness scores. The classification algorithm can achieve up to 0.6 correlation with average human ratings. This demonstrates that facial beauty can be learned by algorithms to a moderate extent. One attribute in the classification of facial beauty, which is important for the model, is the smoothness of the skin using an edge detector. A skin containing lots of edges is categorised as a less smooth skin [Eisenthal et al., 2006, Saeed et al., 2022].

As mentioned earlier, some studies have explored that average faces are attractive but claim that faces with certain extreme features, such as extreme sexually dimorphic traits, may be more attractive than average faces [Little et al., 2002]. Other researchers have suggested various conditions which may contribute to facial attractiveness such as neonate features, pleasant expressions, and familiarity [Zebrowitz and Rhodes, 2002].

Gunes et al. focused on the face region and removed unnecessary parts of the picture. The next steps in their process were the detection of the eyes (position on the face) and the calculation of the geometric values. For the measurement of the proportions of the face, they used a tree-based classifier and used different geometric features. The features can be distinguished in [Gunes and Piccardi, 2006b]:

1. Horizontal: based on the inner-eye distance to the face width.

2. Vertical: based on two important aesthetic theories, namely, the golden ratio and the face thirds method [Ricketts, 1982, Koury and Epker, 1992].

The study of Gunes et al. is based on female facial beauty. Other studies have also explored whether the golden ratios apply to male facial beauty. Peseo et al. describe the similarities and differences between the ratios and measurements of female and male facial beauty. Peseo et al. similar to Gunes et al. base their features on the Golden Proportions and Facial Thirds rules, but both add more ratios and criteria. According to Peseo et al. certain proportions will be perceived as more attractive than others, which can be used in facial beauty detection. However, according to Peseo et al. not all ideal proportions will lead to optimal facial aesthetics, because other influences also play a role here [Perseo, 2002].

Chapter 3

Methodology

3.1 Introduction

Given the broad scope of this thesis project, the methodology provided in this chapter is the general methodology for this thesis project. The chapter consists of two parts: the description of the datasets used for the different experiments and the roadmap of the thesis in general. The specific methodologies of the experiments can be found in chapters 6, 4 and 5, for respectively the human annotation, classification task and the native language search.

3.2 Thesis Process Flowchart

The process of the thesis is displayed in Figure 3.1. The first part of the thesis was the data collection and the annotation. The datasets are described in Chapter 3.3. The data was preprocessed and a selection was made of the best images with the same illumination and same poses (frontal facial images). The annotation of facial beauty are included in the datasets, but the annotations of facial glow are not existing. Therefore, 2000 images are manually labelled with either glowing or non-glowing label.

The dataset conducted in the first part of the process is used for the Neural and Siamese Network experiment. In this experiment different models are used in order to automatically classify either facial glow or beauty. The in-depth methodology is described in Chapter 4.

The third step in the process is the native language search experiment, where a new dataset is created with facial glow images from different regions. An in-depth methodology is described in Chapter 5. The aim of this experiment is to investigate the different perceptions of facial glow around the world using a native language search with a native search engine using a VPN connection.

The final step is the annotation experiment, which was a side project for this thesis. In this experiment 29 participants annotated 50 images. The methodology of this experiment is described in Chapter 6.

The intermediate results of the experiment are discussed each week and the remarks are resolved in this final version of the thesis.

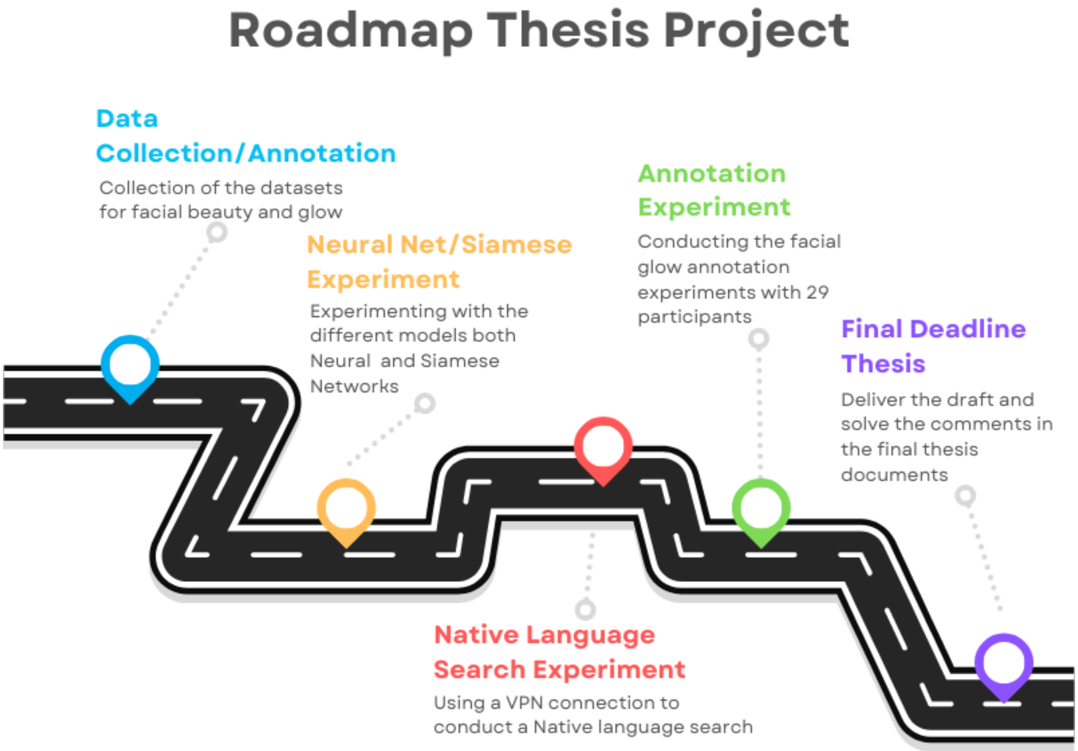

Figure 3.1: Road map of the Thesis Project

3.3 Available datasets for Facial Beauty

One necessary aspect of this research is the availability of large and diverse datasets of facial images annotated with the attribute attractiveness. The datasets are used to train and evaluate the different machine learning models, which can be used to predict facial beauty. Four datasets are explored in this section, including a discussion of their strengths and limitations, namely:

1. UTKface: Contains over 20.000 images with a high level of diversity [Zhang et al., 2017b]
2. Labeled Faces in the Wild (LFW): Contains over 13,000 face images with identity and beauty labels [Huang et al., 2008a].
3. Celebrity Attributes Dataset (CelebA): Contains over 220.000 face images with 40 labels including attractiveness [Liu et al., 2015a].
4. Massively Attributed Aesthetic Dataset (MAAD): Contains over 1 million facial images and has explicit attractive labels [Terhörst et al., 2020].

In Table 3.1 the datasets mentioned above are summarised, including the number of images, the advantages and disadvantages of the dataset.

3.3.1 UTKface Database for Diversity

The first dataset is the UTKface, which has been established in 2017 and the primary use was for age and ethnicity estimation. The dataset contains over 20.000 images and has annotations for age, gender and ethnicity. The large diversity in age and gender is one of the key advantages of the UTKface dataset [Zhang et al., 2017b].

The UTKface dataset has been used in different research studies, including studies on age estimation, face recognition, and facial feature analysis. The dataset is popular due to its applications for fairness of machine learning models, given the diversity of the dataset. This dataset is a valuable source for this thesis project, because the higher amount of inclusive and diverse images of the dataset enables us to understand the relationship between facial beauty/glow and the categories age, gender and ethnicity [Michalsky, 2019, Aruleba and Viriri, 2021, Barrachina, 2022].

3.3.2 LFW Dataset for Beauty

Another well-known dataset is the Labeled Faces in the Wild (LFW) dataset, widely used in face recognition tasks and many other applications in computer vision. The dataset is originally created in 2007 and contains approximately 13.000 facial images including the bounding box and label of the person [Huang et al., 2008a].

The images are collected from the internet, including both posed and unposed images from a range of different ages, genders and ethnicities. A face detector was applied in order to eliminate the false positives. Subsequently, duplicate images were removed and the images were hand labelled by human annotators. Finally, the images were cropped and aligned for the pairwise matching [Huang et al., 2008a].

One of the key features of the LFW dataset is the size and the diversity in the facial images. The dataset not only includes celebrities, politicians, but also regular individuals. Due to the diversity the dataset is a valuable resource for facial recognition and other computer vision tasks [Huang et al., 2008a].

The LFW had been used in various studies, with facial recognition, verification and facial expression analysis. This dataset has also very diverse image features, such as image resolution, lighting conditions and pose, making it possible for researchers to identify the relationship between these image qualities and the performance of the model [Learned-Miller et al., 2016, Karam and Zhu, 2015].

3.3.3 CelebA Dataset for Beauty and Glow

The Celebrity Attribute Dataset is an annotated celebrity dataset with over 220.000 facial images, which was established in 2015. This dataset is known for its large size and worldwide application for face recognition, facial attribute analysis and other computer vision applications [Liu et al., 2015a].

The CelebA only consists of facial images of celebrities, including 40 different annotations including attractive, gender, age, facial hair and eye colour. The images are all collected from the internet, including posed and unposed images. The dataset is very diverse, including people from different fields, such as sports, films and music [Liu et al., 2015a].

The studies on CelebA range from facial attribute analysis and expression to beauty analysis. The dataset is developed to automatically recognise facial attributes, such as age and gender. Lighting condition and image resolution in relationship to facial attribution recognition is also popular with this dataset [Liu et al., 2015a, Wu et al., 2022, Chen et al., 2021].

There has been a lot of critique on the CelebA dataset, due to the accuracy of the attributes in this dataset. The highest agreement for an attribute was received on the label *eyeglasses*, while one of the lowest agreement was reached on the attribute *attractiveness*. This attribute had an accuracy of $< 80\%$ [Wu et al., 2022].

3.3.4 MAAD Dataset for Beauty

Another dataset for facial beauty prediction is the MAAD-face database, which is a novel face annotation database characterised by its large number of high-quality face annotations. In their

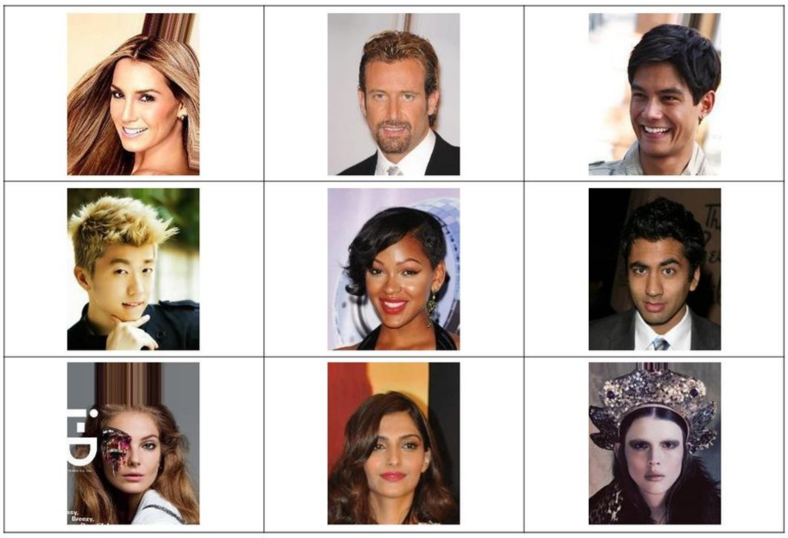

Figure 3.2: Preview of the CelebA database

experiment the Terhorst et al. train a "Massive Attribute Classifier" to extract the attributes of the images. The MAAD-database is an annotation database based on different datasets, including CelebA. For the label *Attractive* the images of CelebA are used, but the annotations are reproduced in order to get a higher agreement score [Terhörst et al., 2020, Parkhi et al., 2015, Cao et al., 2017, Phillips et al., 1997, Huang et al., 2008b, Liu et al., 2015b].

The MAAD has over 3.300.000 facial images from 9.1000 subject with a high variety in pose, age and ethnicity. In total the MAAD-dataset has 47 binary attributes and over 123.900.000 attribute annotations, which makes it one of the biggest annotated datasets [Terhörst et al., 2020].

Due to the increase in performance of machine learning techniques, new methods need to be used to surpass the human performance for recognising age, gender, ethnicity. According to Terhorst et al. this can be done by using soft-biometrics, which can lead to successfully recognise individual and therefore also beauty. This method is according to Terhorst et al. especially powerful when using low quality images and face covering, e.g. masks during the COVID-19 pandemic [Terhörst et al., 2020].

3.3.5 Data Collection for Facial Beauty and Glow

For the purposes of data collection in this research the following steps are involved:

1. The Collection of the `MAAD`-dataset and `CelebA` dataset with human facial images. Both datasets include a diverse range of facial features, including variations in skin tone, facial shape and other characteristics.
2. Both datasets were preprocessed to ensure all images are in the same size and format, remove background noise and irrelevant information.
3. The annotation of facial glow in the `CelebA` dataset. The `CelebA` dataset is chosen, because it has labels for "rosy cheeks", "pale skin" and "heavy make-up" which can be features for the classification of facial glow. A total of 2000 images are manually labelled. This manually labelled facial glow set is used for the training and testing of the algorithms.
4. Exploratory analysis is performed on the `CelebA` dataset to explore correlations between facial beauty and facial glow [Kingma and Dhariwal, 2018].
5. Distinguish a training, validation and testing set is made for both the `MAAD`-dataset and

`CelebA` dataset. The sets are used for the final experiment. The training set is used to train the algorithms, the validation set is used to tune and select the final algorithms, while the test set is used to evaluate the final performance of the algorithms.

Table 3.1: Facial Beauty Datasets

Dataset	No. Images	Pros	Cons
UTKFace	20,000	Large number of images, age, gender, and race labels available	Limited range of attractiveness ratings (1-10)
Labeled Faces in the Wild (LFW)	13,000	Large number of images, beauty labels provided	Limited range of attractiveness ratings, images are not standardized
CelebA	220,000	Large number of images, multiple labels available including *attractiveness*, high-quality images	Label noise, limited range of attractiveness ratings (1-5)
Massively Attributed Aesthetic Dataset (MAAD)	1 million+	Large number of images with explicit *attractive* labels	Limited diversity in subjects and environments

In this thesis project CelebA and the MAAD-dataset labels are used. Moreover, a new dataset for facial glow is created using the CelebA images. Subsequently, the models are then tested on the UTKface dataset in order to evaluate the model performance given the high level of diversity in the UTKface dataset.

Chapter 4

Deep Learning Experiment

4.1 Introduction

In this section the Deep Neural Network and Siamese Network experiment is discussed, including the methodology, results and discussion. The different models are discussed including their best performance on the dataset for both glow and beauty. Both a quantitative and qualitative analysis is provided. The performance metrics given for the different models are the accuracy, precision, recall and f1-score. The findings for facial glow and facial beauty are presented separately due to the distinct character of the two concepts. The different character of the concepts leads to different modelling approaches.

4.2 Applications of Deep Neural Networks

4.2.1 Transfer Learning

Despite the successes of deep learning, it still faces many challenges. These challenges include over-fitting, vanishing gradients problem, model interpretability and computationally extensiveness. Nevertheless, deep learning continues to be an active area of research and many researchers are working on this in order to improve the performance and develop newer, better and faster applications. In computer vision tasks neural networks (NN) have become a very powerful tool to extract the different features in the complex input data [Krizhevsky et al., 2012, Simonyan and Zisserman, 2014, Szegedy et al., 2015a, He et al., 2016a].

The neural networks are trained with back-propagation algorithms in order to minimise the cost function and compared to traditional machine learning techniques, neural networks are much easier in use. Moreover, neural networks also show a better performance compared to handmade feature descriptors. The biggest downside of using neural networks is the computational capacity. The neural networks trained from scratch have a high training time. In order to reduce the training time, a technique called transfer learning can be used. Pre-trained models, which are most often trained on very large datasets, such as the ImageNet, can be fine tuned in order to adapt the neural network on the required dataset [Xu et al., 2018, Bengio, 2012, Yosinski et al., 2014].

Transfer learning can improve the ability of the model to generalise after fine tuning the models, leading to better performance of the classifier on a new target domain. [Yosinski et al., 2014]. For instance, in the study of Donahue et al. pre-trained NN have features containing implicit knowledge, which means that the algorithm performs much better than other state-of-the-art techniques. This illustrates the generality and transferability of NN [Donahue et al., 2013].

In this thesis project the focus is on `Inductive Transfer Learning`. Most algorithms are trained on the `ImageNet` dataset. These algorithms - although having a different data distribution - can be applied to datasets with facial glow or facial beauty. This means that it involves transferring knowledge from a source task to a target task [Torrey and Shavlik, 2010].

In this thesis project the Parameter-based approach is used for the transfer learning. Pre-trained models based on the source domain (ImageNet) are used for this specific target domain (classification of facial beauty and glow) [Torrey and Shavlik, 2010, Pan and Yang, 2010, Farahani et al., 2020, Yang et al., 2020].

The first step is to choose the right pre-trained model. After loading the model with the pre-trained weights a new top layer is made in order to customise the pre-trained model to the target data. This includes the fully connected layer and the output layer. The other layers of the model remain frozen. Subsequently, during the training the engineer chooses which layers are not frozen. The process is iteratively and should be done until the right performance of the model is found. Complex dataset require less frozen layers, while less complex dataset can perform very well with all layers frozen [Torrey and Shavlik, 2010, Pan and Yang, 2010].

The biggest advantages of fine-tuning is that the training of the model takes less time compared to training the model from scratch. Another benefit is that the pre-trained model already has very useful features from much larger dataset, which requires less training on the target dataset. The model can learn the general features and adapt this on the target dataset. Finally, the model also improves its performance, compared to models trained from scratch due to the large amount of source data. This is especially the case when the target and source domain are very similar [Yang et al., 2020, Farahani et al., 2020].

The limitations of fine tuning include computational resources and increased risk of overfitting. First of all, the training time can still require a lot of time given that the entire network needs to be trained on the new task. Additionally, when the target and source domain are very dissimilar, the model is likely to overfit [Vrbančič and Podgorelec, 2020].

4.2.2 Limitations of Convolutional Neural Networks

CNN models are highly effective in many computer vision tasks, such as image classification. However, these models also know limitations. The key limitations are described in this chapter.

The first limitation is the limited receptive field. CNN architectures are designed to process local features of the image. This means that the models have a limited receptive field, which makes it difficult for the model to capture global context or dependencies between distant regions of the image. In short, the CNN models are highly effective at identifying patterns within images. However, the model does not have the ability of effective spatial reasoning tasks [LeCun et al., 2015].

The second limitation is the computational cost of the CNN model. The models are computationally expensive, especially in the case of deep architectures, large images or datasets. This reduces the scalibility and practical use of the models in certain applications. The limited interpretability of the CNN is often referred to as the *black box*. The decision-making process of CNN architectures are often difficult to understand. In some applications interpretability and transparency are important, especially when bias is a high risk, but also in medical diagnosis or legal decision-making processes [Simonyan et al., 2013, LeCun et al., 2015, Gunning and Aha, 2019, Olah et al., 2017].

The third limitation is the dependency of the CNN models on the underlying data. First of all, CNN models are highly sensitive to the quality and quantity of the training data. When the training data is noisy or biased, the performance of the CNN is highly impacted. Moreover, when CNN models require a huge amount of data in order to properly learn the right features and increase the performance. CNN models are prone to overfit on smaller datasets. Overfitting is caused by the lack of the ability to generalise. The algorithm performs well on a training set, but when it faces a unseen dataset, the algorithm results in a poor performance. The reason for overfitting is due to the fact that the algorithm memorises the input and its label, but does not learn the right features. Hence, the lack of the ability to generalise. When applied to other unseen data, the memorised images of the model will not appear, and therefore the model will classify the unseen input wrongly [Goodfellow et al., 2016, Arpit et al., 2017].

Additionally, a CNN has limited ability to handle variability in the data. When the images are high variable, such as changes in lighting, scale or orientation it can highly impact the performance of the model. Techniques such as data augmentation can increase the ability to handle variability in data, but are sometimes not sufficient to address the issue [LeCun et al., 2015, Zhou et al., 2016, Krizhevsky et al., 2017].

The last limitation is the vanishing gradient problem. This problem refers to the loss function regarding the parameters of the neural network. The loss function becomes very small as it propagates backwards through the network during the training. The weights of the model are updated slowly or in the worst case not at all. The result is a sub optimal performance or convergence problems. The risk on the vanishing gradient problem increases as the deepness of the network increases. The gradients become very small in the deep neural networks, due to the many layers the gradients are passed through. This problem can increase the difficulty to train very deep CNN effectively. There are techniques, such as weight initialisation, batch normalisation and residual connection in order to decrease the risk of the vanishing gradient problem [Goodfellow et al., 2016, Zhang and Sabuncu, 2018, Bengio et al., 1994].

4.2.3 Pretrained Models

In this chapter the 6 different pre-trained algorithms used in this thesis project are discussed, namely:

1. VGG
2. ResNet
3. InceptionNet
4. EfficientNet
5. XceptionNet
6. DenseNet

After AlexNet was invented many other CNN models followed. One of the best performing models is the VGG algorithm created by Simonyan et al., which is considered as the most capable models for object-detection due to its simplicity and effectiveness [Simonyan and Zisserman, 2014]. There are two types of VGG models: VGG16 and VGG19. The difference between the two models is the amount of layers (either 16 layers or 19 layers) [Ahadit and Jatoth, 2022, Thareja et al., 2022, Cao et al., 2020, Krizhevsky et al., 2012].

The most common response on the AlexNet was decreasing the window sizes and strides in the first convolutional layer. The VGG algorithm however uses a different aspect, namely the depth. The VGG has the following important differences compared to the AlexNet, namely:

1. VGG is using the smallest receptive field, while AlexNet is using a receptive field of 11x11 with a stride of 4. The VGG is moreover also using three ReLU units, which makes the function more discriminative. Due to this small receptive field, the VGG also has fewer parameters compared to AlexNet's parameters.
2. The convolutional layers of the VGG have the size of 1x1 which makes the decision function non-linear without changing the receptive fields [Alom et al., 2018].
3. The small-size convolution filters result in a high number of weight layers [Alippi et al., 2018].

ResNet-50 (Residual Network-50) is a CNN developed by Kaiming He et al [He et al., 2016b]. The ResNet is a family of models, containing models with the same structure and different depths. In the ResNet-50 the scientists introduced the usage of residual connections, allowing training very deep neural networks while improving the accuracy and gradient flow.

Each of the residual blocks contain multiple convolutional layers. Moreover, the input to each block is combined with the output through a shortcut connection. This improves in bypassing some of the layers in the block. Additionally, it allows the direct flow of the gradient during back-propagation. This helps to minimise the risk of the vanishing gradient problem [He et al., 2015, He et al., 2016b].

The total depth of ResNet-50 is 50 layers, of which are 48 convolutional layers and 2 fully connected layers. The following combinations of convolutional filers are used in these convolutional layers: $1x1$, $3x3$ and $5x5$. Moreover, the model adapts the batch normalisation and the ReLU activation function. This improves the stability and the performance of the model [He et al., 2016b].

The InceptionNet model is also known as GoogLeNet. This CNN is introduced by Christian Szegedy et al. in 2014. This architecture is different from the previously described CNN models due to its use of inception modules. The inception modules allow the model to efficiently extract features at multiple scales. Resulting in the ability to achieve high accuracy on different classification tasks, such as image recognition including facial beauty [Szegedy et al., 2017, Szegedy et al., 2015b].

The logic behind the InceptionNet is to use multiple filter sizes parallel in each layer to extract features at different scales. The output is concatenated to form a multi-scale representation. The inception modules are designed to capture both local and global features [Szegedy et al., 2015b].

Firstly, the inception models, which are the building blocks of the whole architecture, consist of multiple parallel convolutional layers with different filter sizes. The filter sizes are 1x1, 3x3 and 5x5. Followed by a max pooling layer and the concatenation of the output of each inception model. This structure enables the model to capture features at different scales and resolution and makes it useful for the classification of facial beauty and glow due to the fact that the model is able to capture both fine-grained and global contextual information [Szegedy et al., 2017].

Another example of a CNN is the EfficientNet proposed by Tan et al. The popularity of this model can be assigned to its superior performance compared to other CNN with the same amount of parameters. This makes it a well-suited model for various computer vision tasks, such as the classification of facial beauty and glow [Tan and Le, 2019, Tan and Le, 2021].

EfficientNet is based on model scaling. Model scaling involves uniformly scaling the dept, width and the resolution of the CNN, making it possible for the model to achieve good performance. Larger models tend to have better accuracy, but the downside is that the computationally costs are very high. Smaller models have lower computationally costs, but have lower accuracy. EfficientNet aims to strike a balance between the model size and the performance by using a compound scaling method. This method scales the factor for each dimension, which are the depth, width and resolution of the network [Tan and Le, 2021].

There are many types of the EfficientNet, starting with a baseline model named EfficientNet-B0. this is a starting point and other models are scaled up based on this baseline method, making the EfficientNet a family of models. The models are denoted as EfficientNet-B1, EfficientNet-B2 and so on until EfficientNet-B7. All models are increasing in model size. The scaling factors for each dimension are determined by the coefficient ϕ. The ϕ coefficient can be adjusted to control the size of the model. A larger ϕ value leads to a larger model with more parameters, while a smaller ϕ value leads to a smaller model with lower accuracy. The baseline model EfficientNet-B0 has the smallest ϕ value [Tan and Le, 2021, Gan et al., 2022].

The EfficientNet also includes other techniques in order to improve the performance of the model. One of them is the usage of a compound scaling method for the backbone and the feature extraction of the model. Another technique is the incorporation of depth wise combined with point wise convolutions to reduce the computationally cost. The model also introduces the use of swish activation function for better non-linearity calculation. These optimisation techniques allows the EfficientNet to achieve better state-of-the-art performance compared to other models with similar size and higher computationally costs [Tan and Le, 2021, Tan and Le, 2019, Gan et al., 2022, Kartal and Polat, 2022].

XceptionNet is the abbreviation of *Extreme Inception*, which is a CNN designed by François Chollet in 2016. This is a variant of the InceptionNet of Szegedy et al.

The XceptionNet is based on the logic of extreme feature extraction using depthwise separable convolutions. The standard convolutional layers are replaced by two different layers, namely:

1. Depthwise Convolution Layer
2. Pointwise Convolution Layer

The depth wise convolution is an independent layer which performs the convolution for each input channel, while the Point wise convolution is applying a 1x1 filter to combine the output of the depth wise convolution across different channels. The distinction between spatial and channel-wise convolution makes it able for the model to be more computational efficient, due to the reduction of the number of parameters [Chollet, 2017, Wu et al., 2020, Chollet, 2017].

The main difference between the XceptionNet and InceptionNet is the technique of convolutions. InceptionNet is using Inception modules with different filter sizes and concatenates the output. XceptionNet however is using different inception models which are called the depth wise separable convolutions. These are two different parallel operations. The point wise convolution is using a 1x1 filter [Szegedy et al., 2015b].

The last CNN discussed in this research is the DenseNet, which is an abbreviation for Densely Connected Convolutional Networks. This model has been very popular in research due to its unique design which enables features reuse and efficient gradient flow. The DenseNet is characterised by several dense connection in between the layers. Each layer receives input from all the previous convolutional layer. This connection technique is a big difference with traditional CNN models, which connect the layers sequentially [Huang et al., 2017, Zhu and Newsam, 2017].

The multiple dense blocks which are stacked together to form the entire structure of the algorithm. Each dense block has several convolutional layers connected to the previous layer. The dense connections, instead of the sequential connections, allow the model to flow direct information and reuse the features efficiently. These techniques enables higher accuracy and a smaller trainable parameter size [Zhu and Newsam, 2017].

In between the dense blocks the DenseNet uses transition layers. In this layers the batch normalisation, 1x1 convolution and an average pooling is included. The transition layer of the DenseNet reduces the spatial dimension of the feature maps, but also the number of channels. This reduction controls the model size and decreases the computational complexity [Huang et al., 2017].

The biggest advantage of using DenseNet is decreasing the risk of the vanishing gradient problem. The dense connections make it possible to let the gradients flow efficiently through the network, allowing for a better optimisation for the training of this deep neural network. Additionally, the DenseNet also requires less parameters compared to other CNN, as the features are reused effectively [Huang et al., 2017, Zhang et al., 2019].

4.2.4 Siamese Networks

A special type of neural network architecture is a Siamese network. This network is designed to learn the similarity metric between pairs of input, instead of the probability to a class with only one input. The structure of the Siamese Network is shown on Figure 4.1. In this figure two identical sub-networks are shown sharing a similar architecture and weights, hence the name *Siamese*. The input to the model are processed individually, and the outputs are compared to compute the similarity score between them [Melekhov et al., 2016].

A similarity metric between pairs of faces can be very beneficial for facial beauty and glow classification. Siamese networks are performing well on facial verification and facial recognition, making it suitable for facial beauty and glow classification due to the fact that it can handle variations in facial features such as pose, lighting and expression.

What the architecture is of each of the individual CNN models inside the Siamese network depends on the base model used in the Siamese Network. This means that the models as described in Section 4.2.3 can be used in a Siamese network. The input to the network is a pair of images, which are passed to two identical sub-networks, being either one of the pre-trained CNN architectures. The output of each sub-network is a feature vector that encodes the input image and the similarity score between the two images. These are computer by comparing the feature vectors.

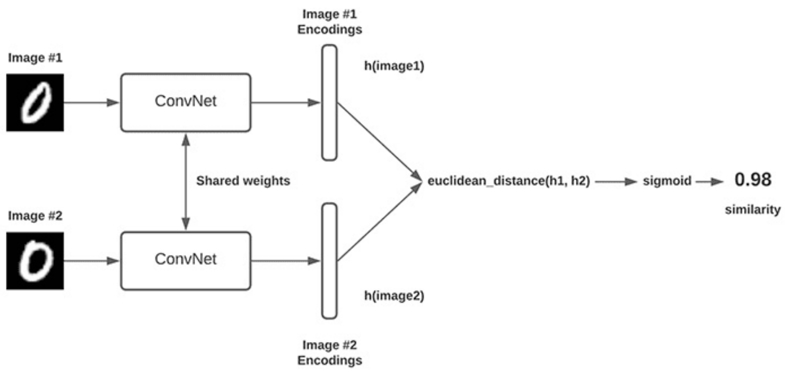

Figure 4.1: The architecture of Siamese Network

During the training the Siamese network learns to optimise a loss function that penalises dissimilar pairs of images. It also encourages the similar pairs to have a high similarity score. The loss function can be adjusted to the specific target domain [Heidari and Fouladi-Ghaleh, 2020].

The advantage of a Siamese network is that it learns to compare pairs of facial images and generates a similarity score based on the facial features. This can be used to classify facial beauty and glow. The ability to compare is especially useful in the case when there is no clear-cut decision boundary between beauty and non-beauty (or glowing and non-glowing). The model can detect subtle differences in facial features and create a similarity score that reflects to a certain degree of beauty or glow [Melekhov et al., 2016, Heidari and Fouladi-Ghaleh, 2020].

Because the Siamese network is using a base-algorithm, this base algorithm can be a pre-trained model. the pre-trained models are very efficient due to less training time and can be fine-tuned on a smaller dataset, which is the case for facial glow. This process saves time and computational resources, but also maintains the high performance [Heidari and Fouladi-Ghaleh, 2020, Yu et al., 2020].

In Table 4.1 an overview with the model, advantages and disadvantages of all models previously discussed are shown. In this table the number of layers and trained parameters are also shown. The VGG-models do not have many layers, but still have a relatively large number of trained parameters compared to other models.

4.2.5 Hyper-parameter Tuning and Settings

The models are multiple times trained with different hyperparameter settings. Next to the data augmentation, multiple hyperparameters were trained in order to prevent the model from overfitting. Given the weak label of *attractive* and *glowing*, the models needed more care and more hyperparameter training in order to make the model learn better and therefore improve the performance. The following parameters are tuned in the models:

1. Momentum Parameter: the SGD optimisation algorithms including different adaptive learning rates were used in order to accelerate the convergence of the optimisation process by incorporating information from previous gradient updates. The goal of momementum is to add a fraction of the previous weights of the gradient update to the current update. This helps the optimiser to overcome local minima, accelerate convergence and reduce oscillations in the optimisation process.
2. Multiple Dense Layers: due to the complex task, the algorithm is prone to overfit very easily very fast. Therefore, multiple dense layers with different parameters are implemented. For each model different settings are used in order to prevent the overfitting and ensure stable results.
3. Different Optimisation Algorithms: different optimisation algorithms are used in this thesis in order to get the best performance for the model. The optimisations *RmsProp*, *Adam* and *SGD* are used. The latter two optimisation algorithms include an adaptive learning rate.
4. Regularisation methods: multiple regularisation methods, such as the L2 and L1 are used in order to get better performance and prevent the overfitting. Not all models performed well with these regularisation methods.
5. Learning Rate: determines how quickly the algorithm learns. When the learning rate is too low, the model may take longer to converge to a good solution, while a higher learning rate can cause the model to not converge to an optimal solution and fail to converge.
6. Batch sizes: different batch sizes are used in order to converge to a local minimum. The best batch size for each of the models was set on 32. When the batch size was reduced or increased, the model did not manage to learn the right features and started to lose performance.
7. Epochs: different epochs are used in order to investigate what the best training time was for the model. The epoch size is set on 50.
8. Fine-tuning: the pretrained models used were trained on the biggest image dataset in the world, namely the ImageNet. Different fine-tuning methods were used, where several layers were frozen, in order to see what the best settings were for the classification at hand.

The best results are showed in this thesis, and in case of significant difference multiple results are shown. The hyperparameters of the models are carefully selected using multiple training trials and `Gridsearch` in order to find the optimal settings for the classification task of facial beauty and glow.

Model	Pros	Cons	Layers	Trained Parameters
VGG16/VGG19	The model architecture is simple and elegant, making it easy to understand and implement it. Has high performance on various range of tasks.	The model is due to the high number of trained parameters computational expensive. The model is prone to overfit on small datasets	16/19	138M/143M
ResNet50	The model uses residual connection to improve the gradient flow and decrease the risk of the vanishing gradient. Has a simple architecture and is easy to train. The model has high accuracy on diverse tasks.	Depending on which model, the computational time is high.	50	23.5M
InceptionNetV3	Multiple filter sizes are used in parallel to capture features at different scales. The computational cost is reduced with the 1x1 convolutions and factorisation.	The model can suffer from an information bottleneck and the vanishing gradient problem	48	24M
EfficientNet-B7	With relatively few parameters and computational resources the highest performing model. Scales well to larger and smaller versions. The model uses a compound scaling method to balance accuracy and efficiency.	The model takes more training time compared to other models. The model is not as well-established or widely used as other models.	813	66M
XceptionNet	The model uses depth wise separable convolution in order to improve efficiency and reducing the number of parameters. With the few parameters used by this model, the model has a high accuracy.	Requires more training data to achieve high performance compared to other models.	71	22.8M
DenseNet-201	High parameter efficiency, due to the fact that the model is reusing feature maps from previous layers. Deep CNN models can be used without encountering the vanishing gradient problem.	The model is computationally expensive due to the dense connectivity.	201	20M
Siamese Network	This architecture learns the similarity score between image pairs. Ability to handle variations in facial features such as pose, lighting and expression. Can be fine-tuned on pre-trained models.	Requires image pairs for training and testing. Requires more training data to achieve high performance.	Depends on network	Depends on network

Table 4.1: Overview of all neural nets and their specifications

4.3 Neural Network Performance for Facial Beauty

4.3.1 Quantitative results for Facial Beauty

In this section the model results for the classification of Facial Beauty are shown. The best results of the experiment trials are displayed. The models are trained with different hyperparameter settings and multiple batches. The performance overview of the models for facial beauty classification are shown in Table 4.2. The best performing model is the VGG19 model, because this model is not only performing as one of the best on the different performance metrics, but also the model is the most stable. The stability of the model is displayed in Figures 4.2 and 4.3. The model is slightly overfitting. However, compared to the other models, especially the larger models such as the DenseNet and deeper ResNet models, this model is the least overfitting model and therefore the model has a higher ability to generalise. The least performing model is the EfficientNet-B7, with an accuracy score of 58%.

Table 4.2: Model Performance Overview Facial Beauty Classification

Model	Optimiser	Dataset	Extra Dense	Accuracy	Recall	Precision	F1-score
VGG16	Adam	CelebA	True	0.74	0.70	0.83	0.76
VGG19	RMSprop	CelebA	True	0.74	0.75	0.75	0.75
ResNet-50	RMSprop	CelebA	True	0.70	0.71	0.71	0.71
ResNet-101	Adam	CelebA	True	0.67	0.67	0.67	0.67
ResNet-152	Adam	CelebA	True	0.66	0.66	0.66	0.66
InceptionNetV3	RMSprop	CelebA	True	0.71	0.72	0.72	0.71
EfficientNet-B7	Adam	CelebA	True	0.58	0.74	0.25	0.37
XceptionNet	RMSprop	CelebA	True	0.74	0.71	0.80	0.75
DenseNet	Adam	CelebA	True	0.72	0.76	0.65	0.70
inception-resnet-v2	Adam	CelebA	True	0.69	0.66	0.78	0.71

The results of the VGG19 model for both optimisers and both labels (either Attractive or Non-Attractive) are shown in Table 4.3. Based on this table we can conclude that the *RMSprop* is a more suitable optimiser for facial beauty classification, because with this optimiser the model performs slightly better for all metrics and for each label. Additionally, using the RMSprop leads to similar scores for both Attractive and Non-Attractive labels. Hence, the results for each label are stable and approximately around 75% accuracy, precision or recall, while the results of other models are more fluctuating and are highly overfitting.

Table 4.3: VGG19 Performance Overview Facial Beauty Classification

Model	Optimiser	Label	Extra Dense	Accuracy	Recall	Precision	F1-score
VGG19	Adam	Non-Attractive	True	0.73	0.74	0.69	0.72
VGG19	Adam	Attractive	True	0.73	0.71	0.76	0.73
VGG19	RMSprop	Non-Attractive	True	0.74	0.74	0.75	0.75
VGG19	RMSprop	Attractive	True	0.74	0.75	0.74	0.74

In Figure 4.2 the loss development of the VGG19 model is shown. In order to prevent overfitting two extra dense layers are used with a double dropout as a top model. Based on Figure 4.2 we can conclude that the loss function is slowly developing over time, with some peaks in the graph. The loss function starts low, but is slightly increasing over the number of epochs. The increase of the loss can be due to various reasons, but one of the most probable reasons for this particular dataset is the high amount of noise in the dataset. The data contains outliers and mislabelled images, which the CelebA is suffering from, which creates the slightly increase of the loss function. However, the models are only trained on a subset of the CelebA dataset for computational efficiency. Training the model on more data could also improve the loss function. Moreover, using a data augmentation technique for each trail the subset of the images can change, and therefore the model results can slightly differ. The model is trained multiple times, and for each trial the model reaches a stable accuracy of 75%. However, the loss over time does change for some subsets, resulting in a steep line for the loss function.

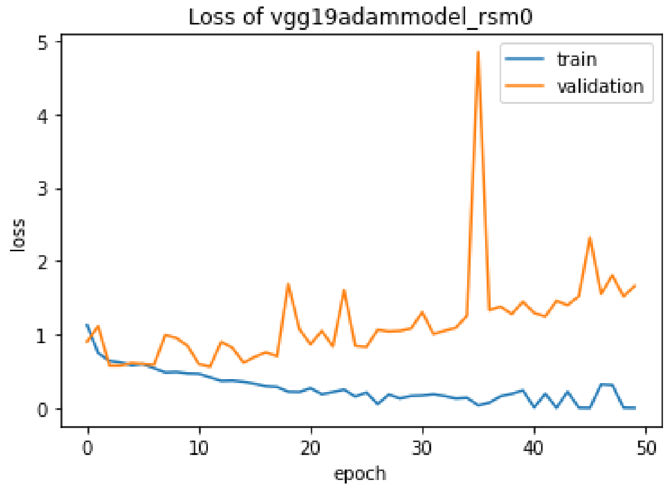

Figure 4.2: The Loss per Epoch for Facial Beauty Classification of the VGG19

In Figure 4.3 the accuracy over the epochs is shown. The accuracy is slowly increasing, with a some outliers, especially at epoch 35. For this figure the same considerations as mentioned above hold. The task at hand is very complex, given the high level of subjectivity. Therefore, the convergence and learning patterns is difficult for the model. Nevertheless, the model results in a stable accuracy of approximately 75% for all labels for female faces, which is a good performance given the complexity and subjectivity of the task.

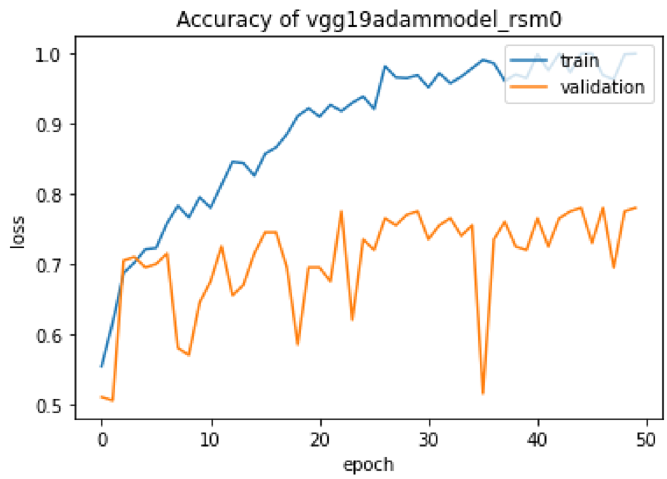

Figure 4.3: The Accuracy per Epoch for Facial Beauty Classification of the VGG19

In Figure 4.4 the confusion matrix of the VGG19 model is shown. The stability of the model is also displayed in this figure, given the fact that the wrongly and correctly classified images are the same for each label. For instance for the set of the Non-Attractive images, only 25 images are wrongly classified as Attractive.

4.3.2 Qualitative results for Facial Beauty

The best algorithm, namely the VGG19, is also tested on the UTKface dataset for the qualitative results. The Neural Networks are often known for their black boxes and therefore making it difficult

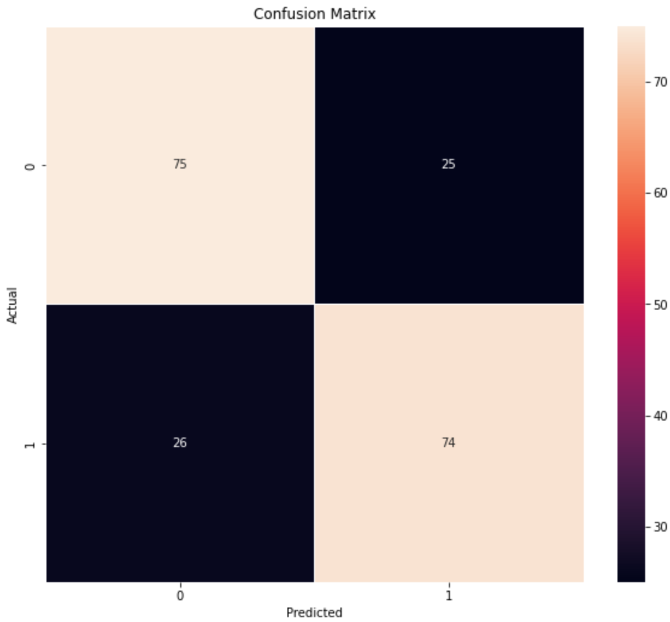

Figure 4.4: The Confusion Matrix for Facial Beauty Classification of the VGG19

to understand what the algorithm is doing and what features are important in the decision-making process.

In Figure 4.5 the most attractive facial images in the UTKface dataset are shown. These images are selected based on the predictions of the Attractive algorithm. In Figure 4.6 the least attractive images are shown.

It remains difficult to understand what the neural networks are doing in their decision making process. However, we can conclude based on the top 10 most and least attractive images, that age is playing a significant role in the decision-making process. Most children facial images are labelled as non-attractive.

Moreover, side facial images are also often labelled as less attractive. All top 10 most attractive images are frontal images. Additionally, the ethnicity of the subjects on the images is fairly mixed. The most attractive facial images still contain a majority of Caucasian faces, but also facial images with different skin tones.

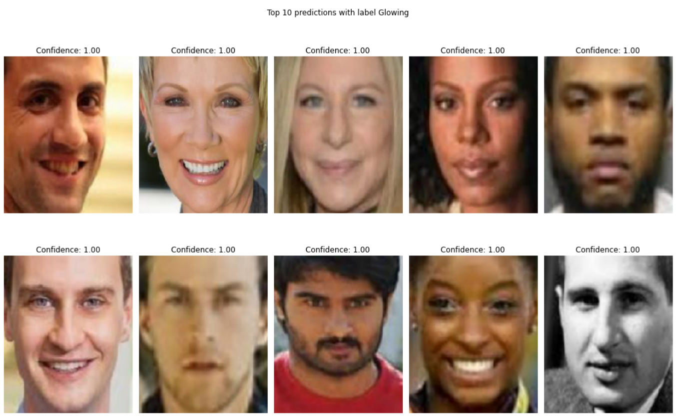

Figure 4.5: The most attractive faces in the UTKFace dataset

Another potential feature can be a smiling face contributing to a higher level of attractiveness. This is also interesting, because smiling faces are often associated with attractiveness. Whether the facial images in the top 10 least attractive images are indeed non-attractive depends on the specific preferences of an individual.

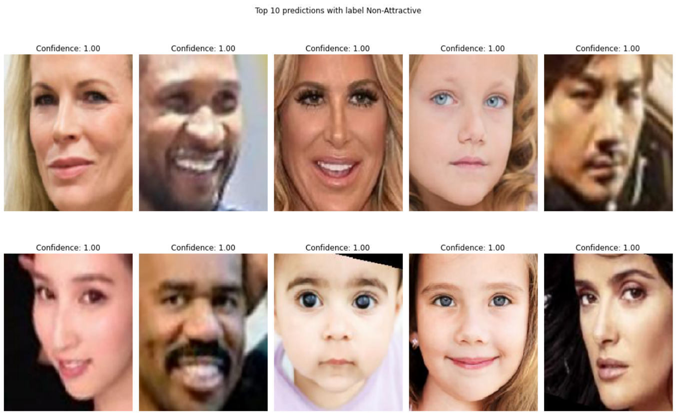

Figure 4.6: The most non-attractive faces in the UTKFace dataset

4.4 Neural Network Performance for Facial Glow

4.4.1 Quantitative results for Facial Glow

In this section the model performance for the classification of facial glow is discussed. The overview of the performance of the models is shown in Table 4.4. The best performing model on this task is VGG19 algorithm with an accuracy of 84%. Similar to the facial beauty classification the VGG19 is the most stable model, giving a similar score for each of the performance metrics. The models are trained with different hyperparameters in order to get the most optimal model for the classification of facial glow.

Table 4.4: Model Performance Overview Facial Glow Classification

Model	Optimiser	Extra Dense	Accuracy	Recall	Precision	F1-score
VGG16	RSMprop	True	0.75	0.75	0.75	0.75
VGG19	RMSprop	True	0.84	0.84	0.84	0.84
ResNet-50	RMSprop	True	0.74	0.77	0.75	0.74
ResNet-101	RMSprop	True	0.75	0.74	0.78	0.76
ResNet-152	RMSprop	True	0.79	0.75	0.86	0.80
InceptionNetV3	RMSprop	True	0.73	0.73	0.74	0.73
EfficientNet-B7	RMSprop	True	0.54	0.52	0.92	0.66
XceptionNet	RMSprop	True	0.72	0.69	0.73	0.71
DenseNet	Adam	True	0.83	0.79	0.89	0.84
inception-resnet-v2	Optimiser	True	0.76	0.74	0.79	0.76

The training loss per epoch is shown in Figure 4.7, based on this image we can conclude a similar phenomenon as with facial beauty classification. The loss function slowly decreases, but after a certain amount of epochs the loss starts to increase. The most probable cause for the increase of the loss function is the high amount of noise and variation in the dataset. Especially images which are close to the boundaries of glow are wrongly classified (similarly for non-glowing images). However, the models are only trained on a small number of images, namely 1600 images as training set, for computational efficiency and due to the only available data. Therefore, training the model on more data could also improve the loss function.

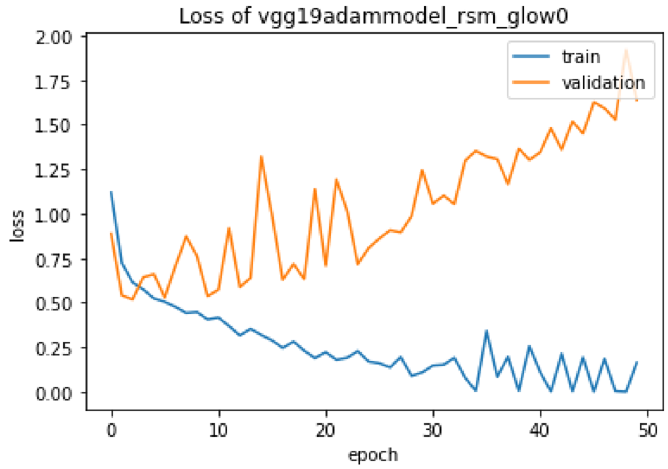

Figure 4.7: The Loss per Epoch for Facial Glow Classification of the VGG19

In Figure 4.8 the accuracy development of the VGG19 model is shown. The accuracy is slowly increasing, with some outliers, especially between epoch 1 to 20. Hereafter, the model has a stable accuracy. For this figure the same considerations as mentioned above hold. Similar to the task of facial beauty classification, the classification of facial glow is very complex due to the high level of subjectivity. The convergence and extracting learning patterns remains difficult for the model. Nevertheless, the model is performing better than the model for facial beauty classification, with a stable accuracy of approximately 84% on all female facial images. Given the complexity and subjectivity of the task at hand, the performance of the model can be denoted as good.

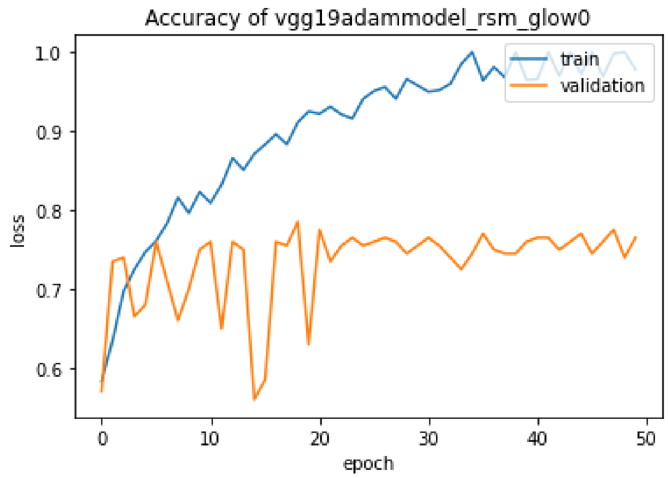

Figure 4.8: The Accuracy per Epoch for Facial Glow Classification of the VGG19

Lastly, in Figure 4.4 the confusion matrix of the VGG19 model is shown. Compared to the facial beauty classification less images are wrongly classified. For glowing faces, only 15 images are classified as non-glowing. These images are often images which are on the boundary of being glowing. Making it difficult for the model to understand the subtle differences.

The results for the classification of facial glow for each label with different optimiser settings are shown in Table 4.5. Using a double dense layer helps the model to prevent overfitting, however the model still tends to overfit, with a training accuracy of 98% and a validation and test accuracy of approximately 84%. Additionally, the optimiser `RMSprop` has a better influence on the performance

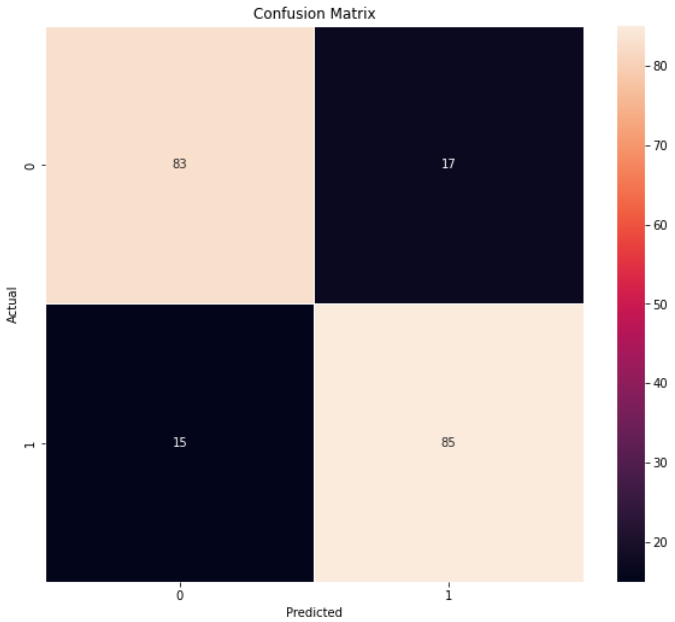

Figure 4.9: The Confusion Matrix for Facial Glow Classification of the VGG19

of the model compared to the `Adam` optimiser. Similar results are shown in the VGG19 model for the classification of facial beauty (Table 4.3).

Table 4.5: VGG19 Performance Overview Facial Glow Classification

Model	Optimiser	Label	Extra Dense	Accuracy	Recall	Precision	F1-score
VGG19	Adam	Non-glowing	True	0.80	0.79	0.80	0.79
VGG19	Adam	Glowing	True	0.80	0.80	0.79	0.79
VGG19	RMSprop	Non-glowing	True	0.84	0.85	0.83	0.84
VGG19	RMSprop	Glowing	True	0.84	0.83	0.85	0.84

4.4.2 Qualitative results for Facial Glow

The Glow algorithm is tested on the UTKface dataset to get qualitative results. In Figure 4.10 the top 10 most glowing faces are shown, with the glow VGG19 prediction, while in Figure 4.11 the top 10 least glowing faces are shown. In Figure 4.10 some accurate predictions are shown (such as the bottom left images), while some other images can be open to discussion. It is also important to note that the algorithm is trained on a female dataset, making it more difficult to detect facial glow in male faces.

Additionally, in the top 10 most glowing images, the ethnicity of the subjects is mixed, but there is a slight tendency to middle aged people. In Figure 4.11 children and elderly are often labelled as non-glowing. The same count for people with a specific skin tone. Additionally, the behaviour of the algorithm seems to satisfy the objective criteria for facial glow set in Chapter 2. The least glowing faces can be associated with less healthy facial images, either based on elasticity of the skin or the illumination of the skin. However, the results are open to discussion given the subjective nature of the task.

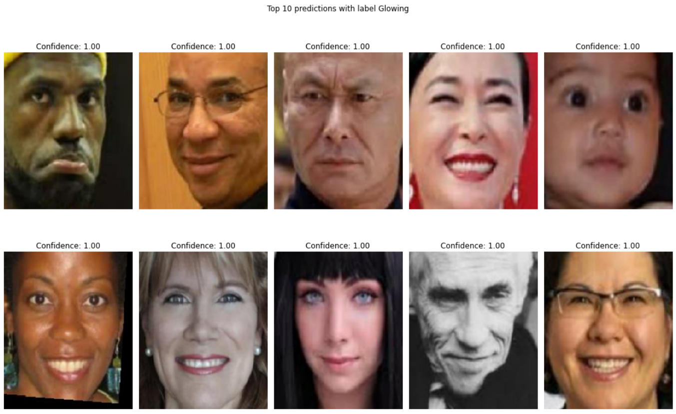

Figure 4.10: The most glowing faces in the UTKFace dataset

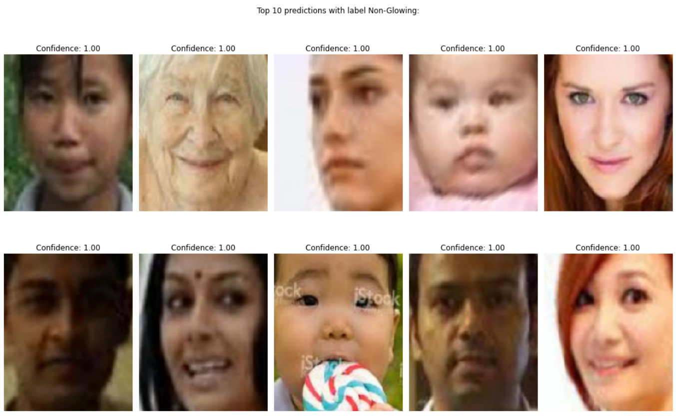

Figure 4.11: The most non-glowing faces in the UTKFace dataset

4.5 Siamese Network Results

In this section the Siamese network model performance are discussed. The Siamese networks have different base models, which are also used in the classification experiment. The models have a pairwise input and the similarity of the images are calculated at the end of the model. If the input images are similar the label is 1, while if the input is dissimilar the label is 0.

4.5.1 Siamese Performance for Facial Beauty

The results of the models for the Siamese architectures of the classification of facial beauty are shown in Table 4.6. As shown in this table, the models are performing significantly worse than the neural networks with a singular input image. The models all have an accuracy around 50%. This means that the models are not learning and do not have the ability to generalise. The best performing model is the `ResNet-101`, which has a stable performance with recall, precision and f1-score higher than 60%.

In Figure 4.12 the training loss over the epochs is shown for the ResNet-101 model. The loss remains very high and does not fluctuate over time. The model does not learn optimally and the weights of the parameters are stuck at a sub-optimal minimum, which leads to poor generalisation and performance of the model. The loss performance over time can be caused by several factors, some of the potential causes are already eliminated, e.g. the learning rate. One cause of this loss development over time can be the learning rate, which can be too high or too low. However,

Table 4.6: Model Performance Overview Facial Beauty Classification

Model	Optimiser	Accuracy	Recall	Precision	F1-score
VGG16	RSMprop	0.49	0.88	0.47	0.62
VGG19	RMSprop	0.53	0.09	0.60	0.16
ResNet-50	RMSprop	0.53	0.36	0.47	0.41
ResNet-101	RMSprop	0.52	0.68	0.64	0.60
ResNet-152	RMSprop	0.51	0.84	0.49	0.62
InceptionNetV3	RMSprop	0.51	0.86	0.49	0.63
EfficientNet-B7	RMSprop	0.55	0.00	0.00	0.00
XceptionNet	RMSprop	0.52	0.49	0.49	0.49
DenseNet	Adam	0.83	0.79	0.89	0.84
inception-resnet-v2	RMSprop	0.55	0.37	0.57	0.45

several experiments are conducted using a different range of learning rates between 0.1 (training very aggressively) and 0.0001. None of the learning rates are improving the performance of the models. Another explanation can be insufficient amount of data. A Siamese network requires a lot data. Increasing the dataset could potentially lead to better performance. Lastly, another probable cause can be mislabelled data or very difficult data. In this task facial glow is very subjective, which means that two images, which are not necessarily similar, can have the same label.

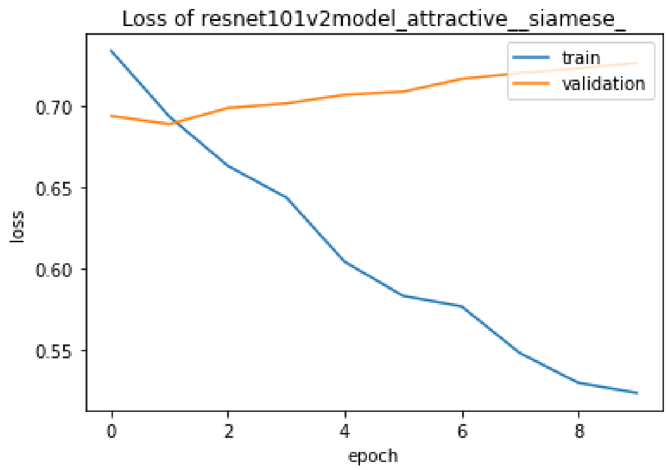

Figure 4.12: The Training Loss per Epoch for Facial Beauty Siamese ResNet Model

In Figure 4.13 the training accuracy of the epochs is shown. The accuracy is not increasing, and remains around 50%. The model is not generalising and leads to a poor performance. One reason for this can be the underlying data, making the model unsuited for this task.

In Figure 4.14 the confusion matrix of the ResNet-101 model is shown. Based on this figure it is clear that the model generalisation is poor and the learning ability of the model is lacking. The number of wrongly classified images is as high as the correctly classified images.

4.5.2 Siamese Performance for Facial Glow

Similarly, the results for the facial glow Siamese network experiment are shown in Table 4.7. Similar results are shown in this table, with all models scoring around 50% on accuracy. The highest performing model with the most table performance is `XceptionNet`. The precision, recall

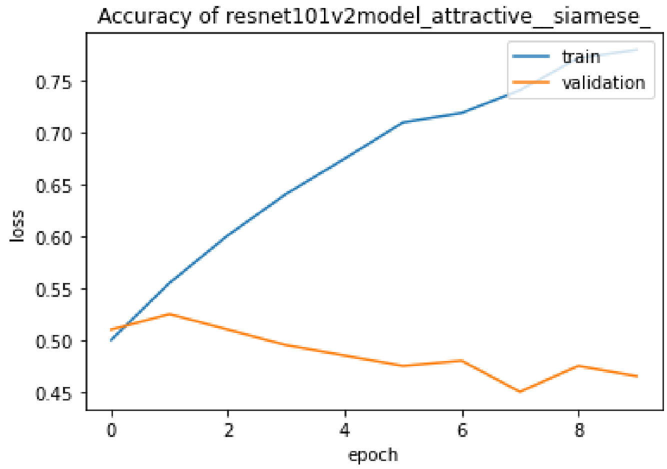

Figure 4.13: The Training Accuracy per Epoch for Facial Beauty Siamese ResNet Model

and f1-score are approximately 60%.

Table 4.7: Model Performance Overview Facial Glow Classification

Model	Optimiser	Accuracy	Recall	Precision	F1-score
VGG16	RSMprop	0.51	0.59	0.55	0.57
VGG19	RMSprop	0.54	0.46	0.54	0.49
ResNet-50	RMSprop	0.57	0.49	0.61	0.55
ResNet-101	RMSprop	0.53	0.67	0.53	0.60
ResNet-152	RMSprop	0.58	0.57	0.56	0.56
InceptionNetV3	RMSprop	0.52	0.57	0.57	0.53
EfficientNet-B7	RMSprop	0.57	-	-	-
XceptionNet	RMSprop	0.54	0.61	0.57	0.59
DenseNet	Adam	0.52	0.14	0.34	0.20
inception-resnet-v2	RMSprop	0.52	0.62	0.30	0.40

Figure 4.15 shows the training loss of the epochs. The loss is immediately very low at the first epoch, which means that the model is not stuck at a local minimum such as the Siamese network for facial beauty.

In Figure 4.16 the training accuracy of the epochs is shown. Similarly as the Siamese model for facial beauty, the model for facial glow is not increasing in accuracy, but remains having a validation accuracy around 50% and training accuracy of 100%, which leads to a heavily overfitting model. The model learning rate is not high and optimised using Gridsearch, therefore an explanation for the model performance is the task at hand. The Siamese model may not be suited for the detection of facial glow, due to the high subjectivity level of the notion.

In Figure 4.17 the confusion matrix of the XceptionNet model is shown. The poor model performance an also be derived from this figure, where the wrong predictions are as high in number as the correct predictions.

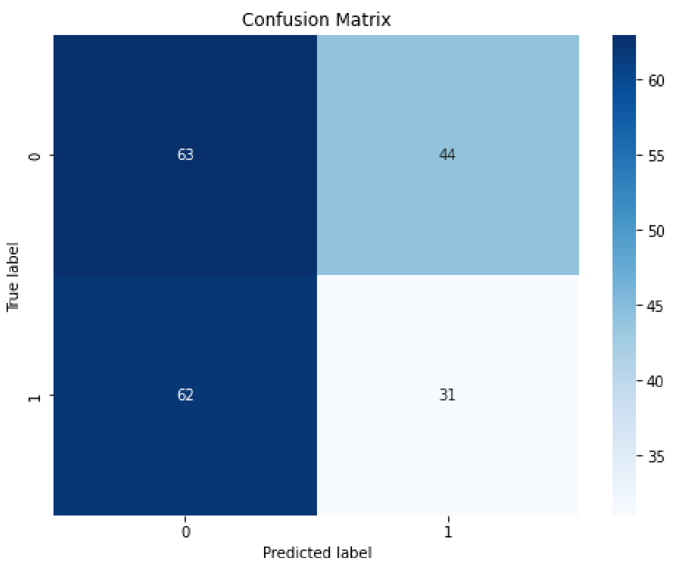

Figure 4.14: The Confusion Matrix for Facial Beauty Siamese ResNet Model

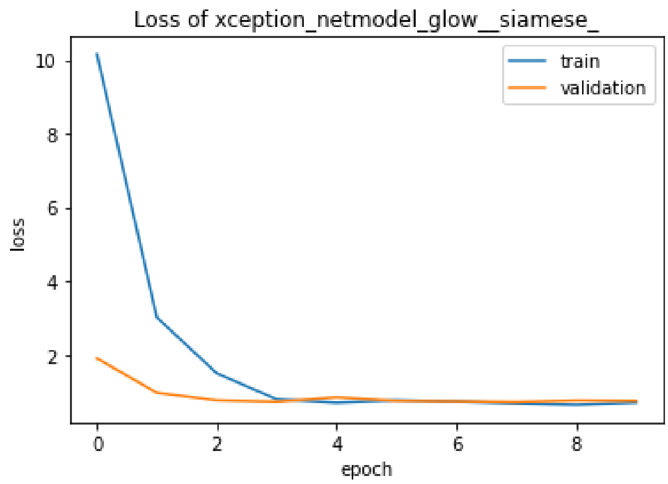

Figure 4.15: The Training Loss per Epoch for Facial Glow Siamese Xception Model

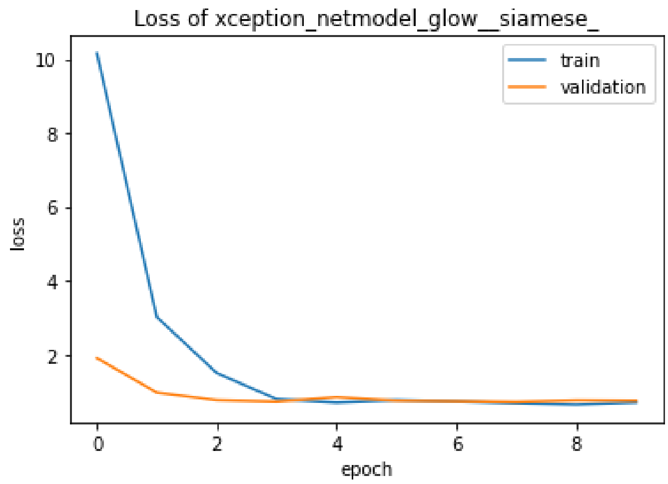

Figure 4.16: The Training Accuracy per Epoch for Facial Glow Siamese Xception Model

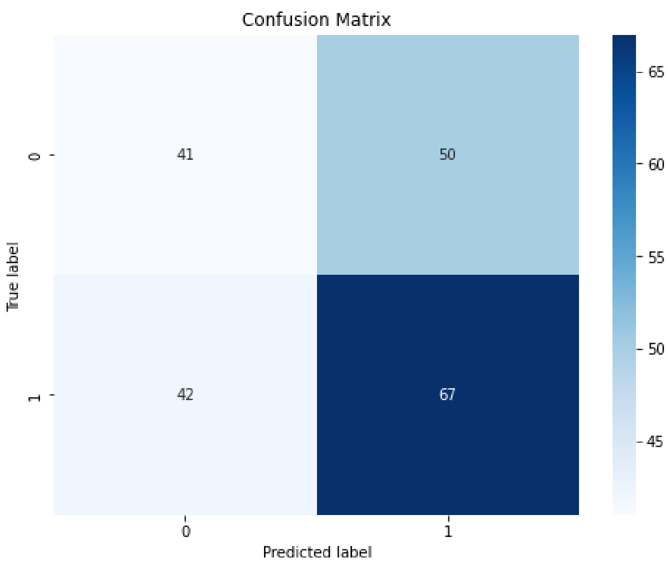

Figure 4.17: The Confusion Matrix for Facial Glow Siamese Xception Model

4.6 Discussion

4.6.1 Limitations of the Neural Networks

One of the biggest limitations is the interpretability of the models used in this experiment. Neural networks are often described as *black boxes* and are difficult to interpret the decision-making process. The lack of interpretability makes it challenging to understand what the network is doing in order to come to a certain conclusion. This is an important limitations because a higher interpretable model makes us better understand the specific features for classification of facial glow.

The interpretability of the model is also important for another reason, namely the underlying bias in the dataset. Neural networks are as good as the training data, which is often associated as the *Trash in, Trash out* principle. When the training data is highly biased or incomplete, the bias is then taken over by the model and reflected on the performance of the network. In the case of the classification task on hand, the risk of bias is high, due to the fact that the CelebA dataset is known for it bias for the label *Attractive*. Moreover, the annotated facial glow dataset is only annotated by one person, which means that the bias risk is higher due to the fact it is only based on the perception of this person. As seen in the experiment with the human annotators, the perception of facial glow can vary over ages.

Additionally, given the high subjectivity level of both facial glow and beauty, the performance of the model reaches a certain point, where it does not increase anymore. This is due to the fact that the classification task at hand is highly subjective and the perception of either facial glow or beauty can vary between different regions, ages and cultures. The model therefore shows signs of overfitting, performing well on the training set, but not as good on unseen data. This means that the modes has not a high ability of generalisation.

Lastly, the models were trained on a small amount of epochs, a total of 50 epochs, due to the lack of computational resources. The image analysis methods used in this study require a significant amount of computational resources due to the large size and complexity of the facial image datasets. However, due to budget constraints, a GPU was not available for use in this research project. As a result, the image processing tasks were performed using a standard CPU, which significantly increased the processing time and limited the size and complexity of the models that could be trained. This limitation may have affected the accuracy of the facial analysis models and the ability to generalise the findings. Additionally, without the use of a GPU, it was not possible to explore more complex neural network architectures or larger batch sizes, which could have further improved the accuracy of the facial analysis models. Therefore, the lack of a GPU in this study represents a potential limitation that should be considered when interpreting the findings.

An alternative approach for the classification of facial glow analysis is the use of traditional computer vision techniques, such as extracting specific features or patterns based on the facial images. These features and patterns can be used to classify the images according to their level of facial glow or beauty. The traditional computer vision techniques can be either local binary patterns of histogram oriented gradients. The features extracted by these techniques can be used to train the classifier with higher interpretability, such as the support vector machine or random forest. These techniques are not only more interpretable, but also computationally faster compared to neural networks and therefore making it more accessible for researchers with limited computational resources. The downside of this method is that these model often do not perform well on very complex and large datasets. Moreover, more time needs to be invested in feature engineering, which can also lead to an increased risk of bias.

4.6.2 The Subjectivity of the Task

Many philosophers, scientists and psychologists have thought about the notion of facial beauty. The concept of facial beauty, but also of facial glow, remain highly subjective. For instance, when taking the experiment results of the human annotators in account, which can be found in chapter 6 we can see that even the human annotators can not find agreement on whether a face is glowing or not glowing.

Looking at the model results, we can conclude that the neural networks are performing significantly better than the Siamese networks. The high subjectivity of the task at hand can be an explanation for these results.

Facial beauty and glow classification involves identifying different facial features that contribute to these notions. These facial features are often highly subtle and require a high degree of precision and detail to capture. A Siamese network is designed for comparing images and contrasting between two input images, while facial beauty and glow classification requires an evaluation of multiple facial features simultaneously.

Neural networks on the other hand are a more appropriate choice for the facial beauty and glow classification. The models are specifically made for image analysis and feature extraction, making them capable of automatically learning the representations of the visual features, but also to capture subtle patterns. These subtle patterns can be very important for the classification, which the Siamese network lack.

Finally, looking at the quantitative results it becomes clear that the neural networks are using age as an important feature for both the classification of glow and beauty. This can be explained due to several factors, such as the underlying data. The training dataset did not have a high variety in ages, and most of the images were facial images of subjects from 20 - 45 years. This makes the data highly unbalanced. Therefore, no conclusions can be derived whether very old or very young facial images are indeed contributing to the perception of facial beauty or glow.

4.6.3 Limitations of the Data

Another limitation is the data collection which can affect the findings of this research project. The facial images were collected based on the CelebA dataset, which is a dataset containing over 200.000 facial images of celebrities. The dataset has a lot of different lightning conditions and variability's in pose.

The CelebA dataset has a limited diversity and is highly biased. The dataset contains images of East Asian or Caucasian people. CelebA is not representative for other ethnic groups and also not a very high quality image database for the generalisation for other populations. Moreover, the dataset also contains limited age variation, making it difficult to use the models on a higher or younger age range.

The dataset is collected from online sources and represents Western celebrities more often than non-Western celebrities. Moreover, the dataset also represents people with a higher visibility or social status more often than others, such as models or actors. Ordinary people are not represented in this dataset, but also other celebrity roles such as top athletes are also not represented well enough by the CelebA dataset.

The MAAD dataset contains the CelebA dataset for the feature *Attractive*, which means that the same limitations hold for this dataset as well. For further research it would be necessary to use a more diverse dataset to decrease the risk of bias. Moreover, the CelebA dataset is also known for its bias in the feature *Attractive*, making the least correlated feature *Male*. Therefore for other studies the recommendation is to use a task specific dataset. An example of this could be creating a dataset with two similar facial images, with one image where the product is applied, while on the second image the product is not applied. The use of task specific images can help the performance of the model in the case of the classification of beauty for a specific product (or the beauty industry).

A subset of the CelebA dataset is also used for annotating glow labels. This means that the limitations for the CelebA dataset also hold for the glow annotated dataset, but several additional limitations can be added. First of all, the annotation is done by one person for 2000 images. Due to the low number of annotators the risk for bias is high, because the annotation solely depend on the perception of one person. However, the perception of facial glow can vary over time, age and culture, which means that it is necessary to include more annotators with different backgrounds.

In the second part of this thesis project a small side experiment was done with 29 human annotators. Based on the results we can conclude that there are many different perceptions of *glow*, which leads to inconsistencies of the labelled dataset. Due to the subjective character of facial glow, it is prone to errors in the labelling process. There is no external reference or validation to ensure the accuracy of the model or the labels.

Despite the limitations described above, the datasets still are a valuable source in this research, but it is important to acknowledge the potential limitations of the approach, especially in the light of the results.

Chapter 5

Native Language Search Experiment

5.1 Introduction

The purpose of this experiment is to conduct differences in beauty standards and skincare practices related to facial beauty and glow across different countries. In this experiment a virtual private network (VPN) and native language search is used to study the different perspectives on facial beauty and glow across various countries. The VPN is used to connect to internet servers located in different countries, including Asian and Western. Native languages is used to retrieve information related to facial beauty and glow for each country. Using this method allows us to gather accurate and culturally relevant information on beauty standards and skincare practices. The results of the native language search are discussed in this chapter.

5.2 Methodology

5.2.1 Virtual Private Network connections

To gather representative and accurate facial glow images corresponding to a specific region, a VPN was used in order to connect to local servers in different countries. Search engines, such as Google or Bing, are using the IP address to determine the location of the user and provide tailored search results. Solely using the native language search did not give the required results and still returned Western search results. There were no to slight differences in the search results when only using the native language search. A VPN however is connecting to a different country and returning a IP address based on that location. Therefore, using a VPN gave more accurate results.

The VPN provider used in this experiment is NordVPN, which is one of the most popular VPN providers. Another reason to use NordVPN is the range of servers in different countries offered by them, which makes it feasible in this experiment. Not all countries are available in the NordVPN services. Most of the countries located on the African continent are not available. The countries used in this experiment are displayed in Table 5.1.

The countries displayed in Table 5.1 are chosen based on several factors. First of all, the main purpose of this experiment is to study the differences between facial glow, beauty standards and skincare practices between different cultures and regions. Therefore, several countries are chosen. The most representative countries are either in the Asian or European region. Secondly, another reason for choosing these countries is based on the high internet usage and search engine activities. This method ensured that the search results are representative of each country's beauty and

Europe	Asian	African	Eastern	Pacific	America
Netherlands (without VPN)	Singapore	South Africa	United Arab Emirates	Australia	United States of America
France	Japan		Israel	New Zealand	Canada
Italy	Hong Kong		Turkey		Brazil
Spain	Taiwan				Argentina
Austria	Vietnam				Mexico
Sweden	Thailand				Costa Rica
Portugal	South Korea				Chile
Greece	Malaysia				
Germany					

Table 5.1: NordVPN Countries used in the experiment

skincare culture. Lastly, the countries are also chosen based on what was available in NordVPN. In summary, the countries were selected based on a combination of research needs and practical considerations.

5.2.2 Native Language Search

The native language search has two components. The first component is the usage of the most popular search engine in each country. Most of the countries, such as European, American and Pacific countries, are using Google as the most popular search engine. However, some countries have other popular search engines as shown in Table 5.2. Therefore, to gather information about facial beauty and glow the use of these specific search engines are important. For example. Baidu is the most popular search engine in China and Naver in South Korea. The search in the remaining countries are done in Google, because this was the most popular search engine for these countries. The language settings were set to the language of each specific country.

Country	Search Engine
European Countries	Google
China	Baidu
South Korea	Naver
Remaining Asian Countries	Google
South Africa	Google
Eastern	Google
Pacific	Google
America	Google

Table 5.2: Most popular search engine in each region

The second component of the native language search is using the right keywords and phrases related to facial beauty and glow. The keywords and phrases should be in the local language, without using the native language the search results are not accurate and mostly result in the Western perspective on facial beauty or glow.

We conducted searches using keywords and phrases related to facial beauty and glow in the local language. An overview of the keywords or phrases are shown in Table A.1.

For example, in Singapore, terms were translated in both Mandarin and Malay. In Malay there are different ways to address facial beauty, namely: "Muka cantik" and "kecantikan wajah". Both refer to the concept of facial beauty. However, there are some slight differences in nuance between them. The first one is a more casual phrase, which is often used in everyday conversation, while the latter is more formal and precise. Moreover, "kecantikan wajah" is not only about physical beauty, but a holistic concept of beauty. Other factors such as skin health, facial symmetry and harmony are also included in this definition. "Muka cantik" however is only describing aesthetically pleasing faces, without necessarily taking into account other factors contributing to facial beauty.

Other languages such as Mandarin and Cantonese also have several keywords or phrases for facial beauty and facial glow, ranging from aesthetic beauty to plastic surgery related. Therefore, it remains challenging to use a native language search while lacking experience with the specific language. In this experiment only a few keywords and phrases are used for the native language search (Table A.1). However, based on the context and the intention of the speaker the phrases may not be adequate in every situation.

5.2.3 Native language image collection

During the experiment the VPN was connected to one of the servers of the countries as listed in Table 5.1. Subsequently, a native language search was performed to gather information about either facial beauty or facial glow. Then a web scraper was used to download the images in a separate folder and in parallel create a csv with the image name and the corresponding label.

The search results were reviewed and relevant articles and websites were selected. The selection was done based on their relevance to the research question. Other patterns, such as prevalence of certain skincare products or beauty treatments were retrieved as additional. The challenges encountered throughout the search process are listed below in order to ensure the validity of the findings:

1. Language barriers
2. Difference in search engine
3. Difference in cultural beauty or glow standards and practices

One major challenge was the language barrier. Most of the languages are not known by me, creating a difficulty in accurately identifying the relevant search terms and understanding the content of some of the search results. In order to address this challenge, a dependency on translation tools and consulting native speakers was necessary in order to ensure the accuracy of the search terms and findings.

The second challenge was the three types of search engines used in the experiment. For most countries Google was the most popular search engine and therefore this search engine was used in the experiment. However, in Hong Kong and South Korea other search engines are used, respectively Baidu and Naver. These search engines operate differently from Google and have their own unique algorithms and ranking systems. This made it difficult to compare the search results across different search engines and to capture a representative sample of information. In order to overcome this challenge, a column denoting the search engine was added in the csv in order to document the search method and ensure transparency in the findings.

The last difficulty is the difference in cultural beauty standards and practices across the world. Some countries, such as France and Sweden, have a greater focus on natural beauty, while other countries, such as South Korea and Japan, emphasise the use of cosmetic products or medical procedures. Therefore it is important to consider the cultural context of the search results to prevent broad generalisation.

5.3 Results

5.3.1 General results

Using the VPN and native language search had the following interesting results.

First of all, between Asian and Western countries there were significant differences in beauty standards and skincare practices. In Asian countries, there was a bigger emphasis on pale, porcelain-like complexion and more youthful appearance through cosmetic products or medical procedures.

Western countries however preferred a more natural look with a greater focus on anti-aging skincare.

Secondly, there are also differences in the types of products and treatments that were popular in each country. One of the examples is the popularity of facial medical procedures in Asian countries, while in Europe only light facial medical procedures, such as Botox, are popular. Additionally, when looking at the types of products, the following can be concluded. In Japan and South Korea, there is a strong emphasis on sheet masks and facial masks, while southern European countries, have a greater focus on high-quality moisturisers, anti-aging products and serums.

Lastly, the language barriers, different search engines and using incognito mode in the web browser made a significant impact on the received results. Not using a VPN resulted in only Western biased results due to the Dutch IP address. Only using Google as a search engine and the English language also led to Western biased results. A similar occurrence transpired in situations where the incognito mode was not activated. To summarise, in order to obtain accurate and culturally relevant results, the experiment should satisfy four requirements:

1. Using a VPN
2. Native language search
3. Native search engine
4. Using incognito mode

Taken all the results into account the experiment validates the significant differences in facial beauty and facial glow perspectives across different cultures and regions. The differences are likely closely tied to cultural values and practices. In order to develop beauty products or marketing strategies the cultural context is of high importance to take into account. Only using the Western language and search engines results in biased result. Therefore, it is important to limit this bias by using the requirements given above.

5.3.2 Country Specific Results

When searching for *Glowing face* or *Beautiful face* in the Asian language equivalent, plastic surgeries were more prevalent compared to searches in other countries. In many Asian cultures a certain standard of beauty and plastic surgery is a viable mean of achieving this ideal. Compared to other countries, there is less social stigma attached to plastic surgery in Asian countries, which make is more accessible and acceptable to a wider range of people.

The most common plastic surgery procedures in Asian countries are double eyelid surgery (blepharoplasty), rhinoplasty and jaw reduction. The double eyelid surgery is the most popular medical procedure, which create a upper eyelid in order to make the eyes appear more Western, while the rhinoplasty is a nose reshaping surgery, which is creating a more symmetrical and refined nose. Jaw reductions are also frequently done in Asian countries in order to create a more feminine appearance [Dobke et al., 2006, Seok-Chan, 2013].

Important note here is that plastic surgery trends can vary widely even within the Asia region. Many factors can influence the types of procedure that are popular in a region or culture. Nonetheless, the results of the experiment validate that medical procedures are more common and socially accepted in Asian countries compared to Western, which is the effect of the cultural differences in beauty and glow ideals.

In Japan there is a strong emphasis on skincare products. Many people are following a multi-step routines that includes cleansing, toning, moisturising and using sheet masks. Another trend popular is *beauty supplements*, which can be capsules, pills or powders, that contain collagen, vitamins (especially vitamin C) or other nutrients. The main beauty perception in Japan is achieving a smooth, porcelain-like complexion. Many Japanese women are searching for products that contain skin-brightening ingredients, like vitamin C.

In most Western countries there is a high demand on high-quality skincare products that contain natural and plant-based ingredients. Moreover, Western women, especially Scandinavian and French, prefer minimalist makeup and are focus on achieving healthy and radiant skin by using serums, oils and other skincare products. The goal is to achieve a natural and effortless look with many women abstaining from heavy makeup and embracing natural features, such as the bushy eyebrows.

Women in the United Arab Emirates, focus on beauty standards with a high emphasis on clear, smooth, and fair skin. Skin lightening products are popular and are often used to achieve a lighter complexion. The most used skincare products often contain oils, which can vary from argan oil or rosehip oil in order to moisturise and nourish their skins due to the harsh conditions of the climate. Another popular product is henna, which is used for beauty and skin adornment. When it comes to make up, the style is significantly different from Western and Asian women. Women in the United Arab Emirates tend to favour bold and dramatic, including the winged eyeliner and smokey eyes.

Another country was Singapore, which had many people favouring natural and organic products, such as green tea, aloe vera and honey. Many products, which were seen in the results were free from harmful chemicals and preservatives. The Singaporean perception of beauty was similar to other Asian countries, which contain clear and bright skin and the use of whitening products to reach this goal.

South Africa is one of the most Westernised countries on the African continent, which also becomes clear in the search results. However, some results also showed some less Westernised beauty and glow ideals by embracing the natural hair and skin textures. In terms of beauty products, most South African women favour products that contain SPF due to the climate and sun exposure. Also products with indigenous ingredients are popular, such as rooibos tea and marula oil.

The American continent in terms of general culture can be distinguished in two parts, one for Northern America (such as Canada and the United States of America) and the other is Latin America. In Northern America, the focus of facial beauty and glow perception is on anti-aging products, containing retinol. Moreover, many Northern American women are seeking chemical free products. However, due to the high diversity in Northern America, such as the states, the perception of beauty and glow can differ across different states (such as a more bohemian style in California, while a more Scandinavian style in New York).

Latin America also varies in different cultures or regions. Although there are a lot of varieties in the perception of facial glow and beauty, there are some commonalities across the different countries in Latin America. One of them is a clear and even-toned skin. In order to achieve this looks moisturisers, serums and mask with ingredients consisting of aloe vera (local ingredient) and vitamin C (mostly extracted from honey).

Another preferred look is the bronzed or sun-kissed look, which leads to the usage of self-tanning products or using oily products in order to tan more easily. Moreover, the results also showed some websites advertising medical procedures, including Botox, fillers, chemical peelings and anabolic supplements.

5.3.3 Clustering the images based on Neural Network Features

For the second part of the experiment, the features of the final layer of the best neural network, VGG19 as discussed in Chapter 4, were taken and used in a clustering model. Five clusters were made based on the dataset and the extracted features.

The clusters are a clear distinction on ethnicity, especially between Western and Asian facial images as can be seen in Figure 5.2 and 5.1. In cluster four there is a mix between Caucasian facial images, while the second cluster contains solely on Asian facial images. The Asian female face in the fourth cluster has Westernised features and may therefore be clustered as such.

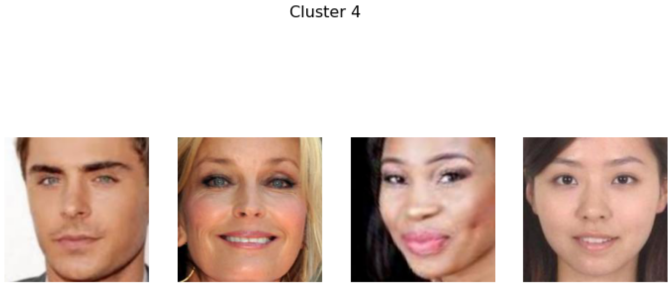

Figure 5.1: The fourth cluster based on the Neural Network features clustering

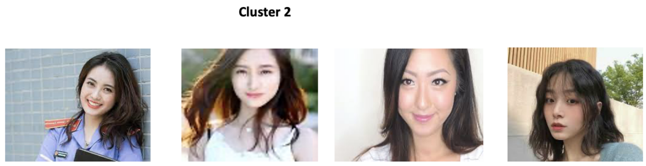

Figure 5.2: The second cluster based on the Neural Network features clustering

Based on the clustered we can conclude that there is a difference in the perception of facial glow around the world. Facial glow in Asian countries is often associated with a very pale skin, as can be seen in Figure 5.2. While facial glow in Western countries is often associated with a sun-kissed skin colour and a healthy look. In Figure 5.3 the first cluster is shown, which are all retouched and highly adjusted images.

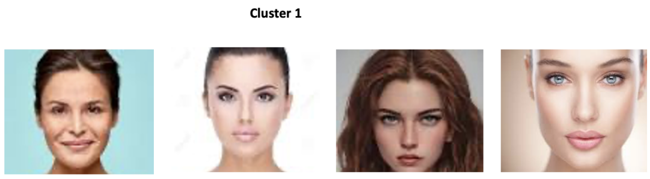

Figure 5.3: The first cluster based on the Neural Network features clustering

5.4 Discussion

This part of the discussion relates to the Native Language Search Experiment. The results of this experiment are discussed in Chapter 5.

5.4.1 Various perception of beauty across different regions

Different countries and regions results in some notable differences in beauty standards, perception of facial beauty and glow, but also in skincare practices. The differences are the results of various factors, including cultural and environmental values, but also historical beauty ideals.

In Asian countries, like Japan and South Korea, the ideal glowing or beautiful face is all related to symmetry and having a fair and porcelain-like complexion. This can already be derived from the translation, where beauty and glow are overlapping. This perception also leads to many skin-brightening and whitening products, but also medical procedures in order to get a more symmetrical face (which can include a double eyelid correction, but also nose surgeries).

Women from Western countries on the other hand have a greater preference on achieving natural and healthy appearances. The focus is more on using high-quality and chemical free skincare products. Medical procedures are a social stigma, therefore not as popular as in Asian countries. This naturalistic and minimalist view can be explained based on the cultural values in Europe considering the historical individualism and self-expression.

Another important factor is the environmental factor, like sun exposure, but also local ingredients. Based on the search results it can be concluded that South African women are using more products which contain SPF and local ingredients such as rooibos tea. The same goes for South American women, where in the results products with aloe vera extract were shown very often, but also Greek search results consisting almost only olive oil products. Countries like Sweden and the Netherlands are less interested in SPF products, due to the less sun exposure and lower temperatures.

To summarise, beauty standards, skincare practices and the perception of facial beauty and glow are shaped by a complex interplay of cultural, environmental, and historical factors. These factors vary widely between different countries and regions. Understanding the differences means gathering deeper insights in the ways beauty ideals are constructed and perpetuated.

5.4.2 Implications for the beauty industry

The findings of this research have important implications for the beauty industry as a whole, but also for individuals. The search results suggest a need for a greater diversity and more inclusive approaches in skincare products and advertisement. Instead of promoting a one-size-fits-all approach to facial beauty and glow, the industry can benefit from recognising and embracing the unique beauty ideals and preferences of different communities. Which can result in developing tailored product lines and advertisement reflecting the cultural and environmental factors that shape beauty standards and skincare practices in each region.

Additionally, the results also makes room for opportunities in cross-cultural exchanges and learning within the beauty industry. Knowledge and expertise sharing from different cultures and region can lead to a more inclusive and diverse approach to facial beauty and glow instead of the narrow and often unrealistic beauty ideals that dominate mainstream media and marketing campaigns.

The results also can affect individuals in highlighting the importance of understanding and accepting diverse beauty and skincare practices. Recognising the unique beauty ideals and skincare needs of each community can benefit individuals in making more informed choices about their own beauty and skincare routines and avoid harmful beauty stereotypes.

Based on the findings above, it can be concluded that there is a need for a more nuances and inclusive approach to facial beauty and glow. It is a necessity for the approach to recognise, acknowledge and highlight the unique beauty ideals and skincare practices of different regions across the world.

5.4.3 Limitations of the experiment

The experiment encountered several challenges and therefore has several limitations. The first limitation is the usage of VPN to access search results from different countries. The sample used in this experiment is limited to the countries provided by NordVPN. Hence, not all regions in the world are represented as well as others. One of the most important regions is Africa, from which only South Africa, which is already Westernised itself, was included in NorthVPN. Therefore, the results of South Africa is not representative for the continent Africa. Moreover, in this experiment it is assumed that each country has its own general perception of facial beauty or glow. However, even in a small country like the Netherlands the perception of beauty or glow can differ amongst cities. Amsterdam is using a more Scandinavian approach, while other cities, e.g. Rotterdam and The Hague, are not. For further research it is necessary to include a larger and more diverse sample of countries and also look at a more granular level, e.g. cities or provinces.

A second limitation is the methodology of using online search results. Some regions are more connected to the internet than others which can lead to not have the full picture of beauty ideals or skincare methods. Moreover, some cultural practices or beauty rituals are not widely discussed online. Therefore using a more in-depth interviews or focus groups can provide a more nuanced understanding.

Additionally, this experiment did only explore the perception of facial beauty and glow for females and not for males. These two groups significantly differ in their beauty and skincare practices, e.g. in general women are using more make-up and have a stronger preference for medical procedures. Hence, the conducted experiment did not explicitly explore the intersection of beauty ideals and issues of gender, which is known to play a significant role in shaping beauty standards and practices. Future research could delve deeper into these issues to provide a more comprehensive understanding of the perception of facial beauty and glow.

Finally, most of the languages are not known to me personally, therefore the experiment was heavily depended on translation tools and native speakers. However, some languages have a lot of nuances for the words *facial beauty* and *facial glow*. As stated in Section 5.2.2 in Singapore different languages are used by the citizens, such as Malay, Chinese and English. In Malay the keywords for *facial beauty* can be either translated as "Muka cantik", but also "kecantikan waja". This same rationale applies not only to Asian languages, but also European languages. Thus, in order to conduct an exhaustive search for the right keywords, proficiency in the respective language is a necessity. Relying solely on translation tools is not adequate enough to guarantee accurate and comprehensive outcomes.

This experiment is just the beginning of a more thorough research in facial beauty and glow. While it has provided some valuable insights in the perception of facial beauty and glow across different regions in the world, there is still much to be explored and studied in this research area.

Chapter 6

Annotation Experiment

6.1 Introduction

In this chapter the side annotation project is discussed. The side experiment was executed with female participants annotations each 50 facial images based on the CelebA dataset. The goal of this experiment is to understand the most important requirements for facial glow.

6.2 Methodology Annotation Experiment

6.2.1 Data Selection and Participants

A subset of 50 images is taken from the CelebA dataset in order to annotate for either glow or non-glow faces. The task given to the annotators was to choose the most glowing image between two displayed female facial images. First of all, only female faces without eye accessories are chosen in this subset. Subsequently, the images with the same illumination are filtered, faces without normal poses were excluded, such as a screaming face or other irregularities. This resulted in consistent images and reduce the influence of extreme images on the outcome of the annotation.

In total 29 participants of different nationalities and ages have annotated the same dataset. In Table 6.1 the different ethnicities of the participants are shown. Ethnicity is not the nationality, but the origin of the participant. The most participants are from the Netherlands.

Before the experiment the conditions for facial glow, as discussed in Chapter 2.5, are explained to the participants. The six requirements of facial glow are discussed with the participants, namely:

1. Luminosity of the skin: the overall radiance of the skin, especially in the T-zone.
2. Smoothness of the skin: the absence of wrinkles and imperfections of the skin.
3. Hydration of the skin: the hydrated appearance of the skin.
4. Evenness in skin-colour: the evenness in colour without any hyperpigmentation of the skin.
5. Elasticity of the skin: the ability of the skin to stretch and return.
6. Visibility of the pores: the absence of blackheads and facial pores.

The participants were provided with clear instructions and guidelines for the annotations, in accordance to the criteria described above. The images are shown pairwise to the participants. Every participant is shown the same set of images.

6.2.2 Fleiss' Kappa Coefficient

As previously mentioned, the annotation of facial glow in facial images involves human participants and influenced by their subjective perceptions. To gather insights in the inter rater agreement for the facial glow feature, the Fleiss kappa statistic is used. This statistic considers the possibility of chance agreement and is commonly used in fields such as psychology and medicine to assess the reliability and agreement among two or more annotators.

The Fleiss kappa statistic, denoted as κ, is calculated using a similar formula to the Cohen kappa statistic. However, instead of comparing two raters, it compares the level of agreement among three or more raters. The equation of the Fleiss' kappa is:

$$\kappa = (P_o - P_e)/(1 - P_e) \tag{6.1}$$

where P_o is the observed agreement among the raters and P_e is the expected agreement due to chance. The data is organised into a contingency table, where each row represents the annotation of a single rater, and the columns represent the different categories being rated. The diagonal sum divided by the total number of ratings yields P_o, while P_e is calculated by multiplying the total number of ratings by the product of the marginal probabilities for each category.

The Fleiss kappa statistic ranges from 0 to 1, with 0 indicating no agreement beyond chance and 1 indicating perfect agreement. Negative values indicate that the raters are in disagreement beyond what would be expected by chance. In order to establish a strong ground truth, a value of $\kappa > 0.8$ is recommended, indicating a high level of agreement among the raters.

In this research project, the Fleiss kappa statistic is applied to the annotations of facial glow by the human participants.

6.3 Exploratory Analysis for Facial Beauty

Before executing the experiment an exploratory analysis was executed on the same 50 images of the CelebA dataset. This exploratory analysis included a correlation matrix, which is shown in Figures 6.1 and 6.2. The top 10 highest and lowest correlated features to the attribute *Attractive* are depicted in these figures. The correlation matrix provides the pairwise correlation between the different features in this dataset.The value -1 indicates a perfect negative correlation, while 1 indicates a perfect positive correlation. This matrix is used to identify patterns and relationship between different features.

In Figure 6.1 the top 10 highest correlated features to the attribute *Attractive* are shown. Based on this graph, we can conclude that almost all highly correlated features are female features, such as `Heavy makeup` or `Wearing Lipstick`. Both features correlate to the label Attractive with a Pearson correlation of 0.48.

Additionally, in Figure 6.2 the top 10 lowest correlated features to the attribute *Attractive* are shown. The lowest correlated feature with the *Attractive* is the feature *Male*, with a Pearson correlation of 1. This means that the dataset is highly biased towards female attractiveness. For this experiment only female faces were used for the classification task. The second lowest correlated features is `Double Chin`, with a Pearson correlation of 0.53. Other important features mentioned in this figure are `Gray Hair, Blurry, Bags Under Eyes and Chubby`. The feature `Gray Hair` is the opposite of the high correlated feature `Young`, due to the fact that younger people often have less gray hair. Moreover, the annotators of the CelebA dataset also favoured a more "healthy" look, given that 3 of the lowest correlated features are associated with physical healthy appearance.

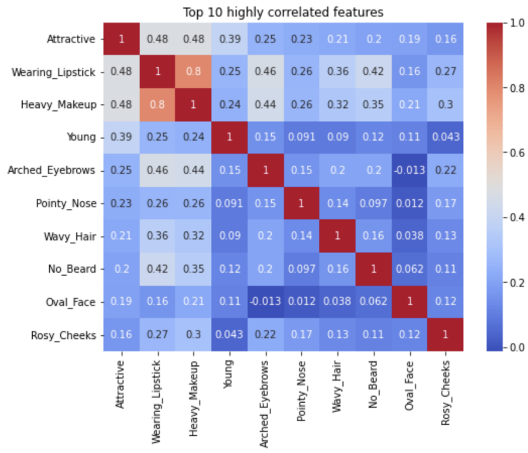

Figure 6.1: Top 10 highest correlated features with Attractive

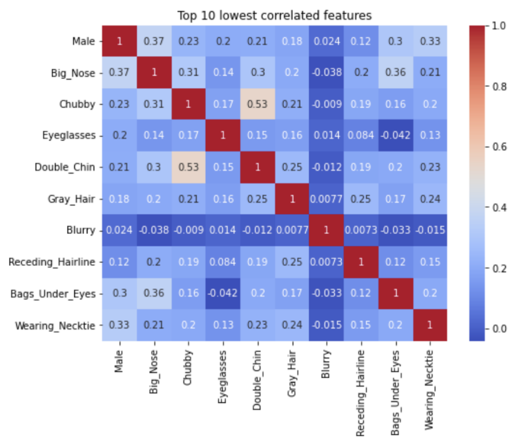

Figure 6.2: Top 10 lowest correlated features with Attractive

6.4 Glowing Face Annotation results

The same image pairs are annotated by female participants. The participants had different backgrounds. The ethnicity of the annotators is shown in Table 6.1. Most participants, 41% of total are Dutch.

In Table 6.2 the distribution of the age ranges are shown of the participants. Most of the participants, 58%, were young adults ranging from the age 25 - 34 years and just graduated of college and were working. Only few people were either younger than 18 or between the age range of 45 - 54 years.

In Table 6.3 the top 10 highest correlations with facial glow are shown. Based on this image we can conclude that there are no high correlation with the pre-annotated labels of the CelebA. This also became clear when asking the post-experiment questions. 78% of the candidates agreed that facial glow did not have a relationship with facial attractiveness. Facial glow was perceived as having a healthy skin, while facial attractiveness included other facial features, which was not solely a healthy skin.

Additionally, a majority of the participants, namely 87%, mentioned that make up was an indicator

Category	Total
Armenian	4
Chinese	1
Dutch	12
Hungarian	1
Kurdish	1
Russian	4
Suriname	4
Ukrainian	1
Romenian	1
Total	29

Table 6.1: Ethnicity distribution by category

Age Category	Total
<18	1
18-24	5
25-34	17
35-44	1
45-54	3
≥ 55	2
Total	29

Table 6.2: Age distribution by category (sorted by age)

for facial glow. However, the participants also mentioned that having heavy make up was not preferred, but an overall neat appearance was. In the correlation matrix shown in 6.3 the attributes *young, heavy makeup, wearing lipstick* are shown as light correlations with the attribute glow. This can be confirmed by the answers on the open questions by the participants.

Another important indicator given by the participants was the overall appearance of health of the subject shown on the figure. In Figure 6.4 the lowest correlations with facial glow are shown. One of the lowest correlations with facial glow is having *gray hair* or signs of overweight, such as *chubby* and *double chin*. These answers were also given by the majority of the participants. Having an overall healthy appearance, such as a normal weight and a skin free of acne and other irregularities was decisive for their choice of the annotation of facial glow.

Moreover, younger participants tend to classify images with darker skin-tone with a higher level of facial glow compared to the older participants. This can be due to the diversity of the younger participants, while the diversity in nationality of the older participants is very little.

The annotations are reviewed with the use of the *Fleiss'* kappa. The interrater agreement for facial glow is low, with an Fleiss' kappa of -0.007. This means that among the participants the agreement for facial glow is lower than given by chance. Hence, whether an image is glowing or not glowing has a lot of disagreement.

However, when participants are grouped by age, the trend of disagreement shifts. The participants under the age of 34 (the lowest three age ranges) show a higher agreement level, a Fleiss' kappa of 0.59. This means that there is a moderate level of agreement amongst the participants. When the older two age groups are grouped, the Fleiss' kappa increases to 0.65. The results show that age might be an important indicator influencing the annotations of facial glow.

Because the number of data points on ethnicities is small, no calculations can be performed based on this subset. Due to the low representation of different ethnicities in the sample size no meaningful conclusions can be drawn.

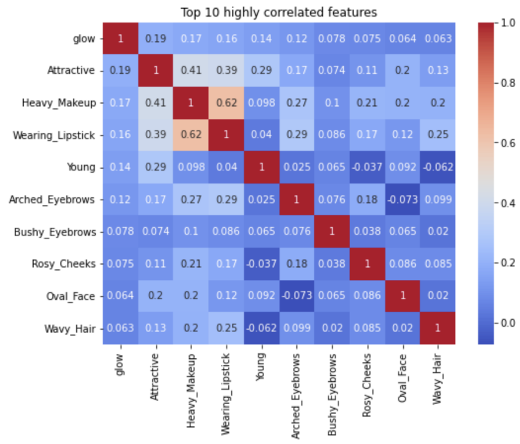

Figure 6.3: Top 10 highest correlated features with Glowing Face

The findings above suggest that the age of the annotator can be an important indicator for the perception of facial glow. Older participants did not perceive slightly wrinkled skin as less glowing, while younger participants did. These results suggest for more targeted marketing campaigns, instead of using only young models promoting facial glow.

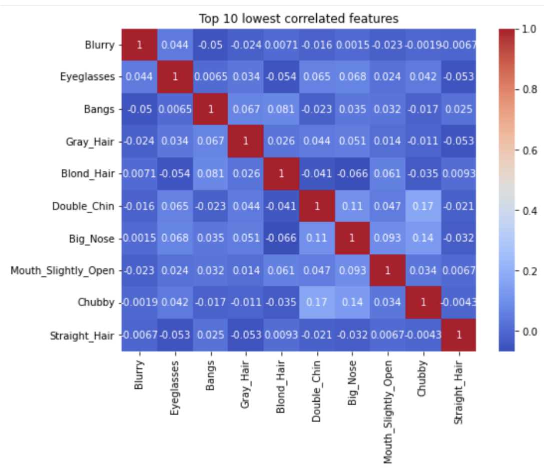

Figure 6.4: Top 10 lowest correlated features with Glowing Face

In Figure 6.5 an example image is shown. Approximately 90% of the annotators agreed that the left image was more glowing compared to the right images. The reason given by the participants for this is the sun tanned colour and the healthy look of the female. Figure 6.6 shows two facial images where the participants scored a low fleiss kappa between the different age groups. Participants within the younger age group (up until 34) had a higher agreement labelling the right female images as more glowing, while older participants (participants from 45 on wards) labelled the left female image as more glowing.

After completing the annotation task of the facial images, the participants were asked four post-experiment questions in order to gain deeper insights in their perception of facial glow. One of the main findings was that most of the participants did not associate facial glow with facial beauty, with some of them perceiving images as facial glow, while not as facial beauty. Younger participants tended to place more importance on facial glow as a component of facial beauty

Figure 6.5: Facial Images With a High Fleiss' Kappa Coefficient

compared to older participants, who focused on the overall appearance.

The images contained only female faces, and for further research the question was asked whether facial glow would be different for male faces. Participants generally agreed that facial glow would be more difficult to perceive in male images due to the facial hair. However, there was no agreement between the participants, whether facial glow was solely an attribute of the female face.

Figure 6.6: Facial Images With a Low Fleiss' Kappa Coefficient

6.5 Discussion

One of the biggest limitations of the annotating experiment is the limited number of human annotators. In total 29 human participants took part of the annotating experiment. These human participants were all female and from different age ranges. However, the majority, 58%, of the participants belonged to the age range 25 - 34, which means that the results are not representative. Moreover, the ethnicities of the annotators were also limited, meaning that nothing can be said on the aggregated level of ethnicity. Therefore, it is important to acknowledge the potential impact of the size of the participants and the reliability of the results. With 29 participants the risk that individual biases or preferences have a disproportionate influence on the overall findings is increased. Increasing the number of participants in future studies can help to ensure more reliable and representative results.

In order to get a better understanding of the perception of facial glow a larger amount of participants are needed with a more diverse background in both age, gender and ethnicity. The diversity of the participants can be used in the analysis to gain deeper understanding of the concept at hand. A diverse group of annotators can reduce the risk of potential biases, but also increase the likelihood to capture a wider range of perspectives related to facial glow.

The annotators were presented 50 images, therefore the number of images also impacts the quality of the results of the experiment. Using 50 images is not representative in order to get a full understanding of the features contributing to facial glow. The findings in this study therefore may not fully capture the complexity and variability of this subjective notion in real-world context.

To improve the quality of the findings, it is necessary to include a higher number of facial images, using different poses and lightning conditions. Moreover, the facial images should also contain people with different skin tones, facial shapes, genders and age in order to capture the full range of indicators influencing the perception of facial glow. Including images with makeup or facial hair (for male facial images) can also lead to different results.

Chapter 7

Discussion

7.1 Introduction

The focus of this master thesis was on the classification of facial beauty and glow. Three experiments have been conducted in order to gain a deeper understanding of the concept of facial beauty and glow and the interaction between these two concepts. These experiments however have certain limitations and considerations, which can be used for further research. The limitations and considerations of the experiments conducted in this master thesis are discussed in this chapter.

7.2 The Power of Perception

Both facial beauty and glow are highly subjective concepts, which makes it difficult to reach an agreement between annotators. Even in the facial glow annotation experiment, some annotators agreed fully on one image, while not on other images. Individual preferences, regional influences and other factors can cause this huge difference between the annotators. The Siamese networks, in Chapter 4, also have a very low performance. The performance of the Siamese Networks can be explained by various factors, such as the model architecture, amount of training data, mislabelled data, but also the subjectivity of the task presented in this thesis.

One of the biggest questions remains is: is facial beauty really objective? And what about facial glow? Based on the performance of the neural network performance we see that the performance of facial glow classification models is significantly higher than the performance of facial beauty classification. In my opinion, this can be explained due to the semi-objective character of facial glow. Facial glow is often associated with healthy appearances and therefore has more objective criteria compared to facial beauty. An skin full of acne will not be considered glowing, while a healthy hydrated skin with a certain amount of make up is considered as glowing. These criteria are easier to grasp and agree on compared to the features of beauty.

Facial beauty also has different requirements based on literature. However, in reality most people do not agree on this particular notion. Person A can be attracted to female faces with bushy eyebrows, while person B associates this features with masculinity and therefore unattractive. Even small changes in the appearances can change the perception of beauty. Therefore, while facial glow can be understood in terms of healthy appearances, facial beauty is stuck in the eye of the beholder.

7.3 Social and Ethical Impact

The classification of facial beauty and glow can have great impact on social and ethical implications and therefore should be examined carefully. First of all, the bias in the dataset can lead to potential discrimination or stereotyping based on the facial appearance. Categorising individuals based on facial glow or beauty can harm beauty ideals and exclude different groups from this perception. Moreover, accepting that there is one ultimate and 'right' ideal of beauty or flow can impact and harm individuals, due to the unrealistic standards. The biases can lead to social hierarchies or impacting people's self-esteem and social interactions. Classification of these kind of sensitive topics should be used with care.

Hence, the potential bias and inaccuracies in facial beauty and glow classification should be acknowledged. The training data used for these experiments were not diverse and representative. Subsequently, it can perpetuate biases and reinforce existing societal inequalities. Therefore it is crucial to have an inclusive dataset with different ethnicities, ages and genders.

However, the question remains if we should want to classify facial beauty and glow. Facial glow should then - taking all ethical implications in consideration - be limited to the objective features such as a healthy appearance, while facial beauty classification may not be ethical at all. Facial beauty can vary over time and regions, but also over one's sexuality. Proposing a standard classification algorithm will introduce a one-size-fits-all approach, which is not maintainable in our society.

7.4 Implications

The implications of this facial glow analysis vary over different fields and industries. First of all, this is one of the first studies which used an automatic classification of facial glow. The use of the methods described in this thesis, can have great implications for the beauty industry, especially in marketing and advertisement. First of all, facial glow is highly associated with health. Almost all images either labelled by the human annotators or found through the native language search, contain images of good skin health and quality and are not necessarily associated with attractiveness. Based on the results we can assume that facial glow is a health indicator and not an attractiveness indicator. The facial glow analysis can also be used in healthcare to indicate the overall health of patients, given the fact that glow and health are highly related.

Moreover, the models trained for the facial glow analysis can be used in order to find the right marketing campaigns for skincare products. A dataset with facial images, containing *applied skincare product* and *not applied skincare product* can be used to see whether the product is increasing the radiance of the face. The development of new products can also improved by using the models to predict whether the product is indeed increasing the radiance.

Secondly, the results of this thesis project also have implications for different research & development departments regarding the production of new products. These results provide valuable insights into the preferences of customers and their perception on beauty and glow. This information can be used to develop new products and improve existing products to align with customer desires and expectations. Understanding which features contribute to the perception of beauty and glow provide guidelines for the development of cosmetic procedures and products, as well as for skincare treatments and digital tools.

7.5 Future Research

For the purpose of further research several methods can be applied. These methods are described in this chapter.

7.5.1 Creating an Inclusive and Large-Scale Dataset

First of all, a new dataset for the classification of both facial glow and beauty is required in order to develop a more accurate and generalised model. Despite the availability of several datasets, such as the MAAD and CelebA, these datasets still deal with an increased risk of bias and the lack of diversity in the sample images. For the classification of facial glow no datasets are available, and therefore there is a need for a new dataset containing multiple features, including the labels *Attractive* and *Glow*, but also other indicators such as *Skin Quality* and *Golden Ratio*.

A larger dataset provides more information and is more appropriate to train robust and accurate neural networks. Moreover, a higher diversity in the dataset can help us understand the differences in the perception of facial glow or beauty between different ethnicities and regions. Various sources can be used to create the dataset, but the most prominent source would be social media platforms, due to the fact that these platforms have the highest amount of "real-world" applications, with a large number of images and a high variety of lighting conditions and populations. The images then can be used on different annotating platforms in order to gain more perspective on the perception of facial glow and beauty.

Moreover, in order to understand the differences in perception between female and male images, it is necessary to include other genders in the dataset. The experiment in this project is done for female facial images, but as the participants in the annotating experiment, Chapter 6, already confirmed the perception of facial glow can differ significantly when using male facial images due to facial hair and different skin texture.

The concepts of facial beauty and glow can differ over regions, cultures, but also across fields. Social media has an alternative perspective on these concepts, compared to plastic surgery or art. Therefore, further analysis is needed to not only distinguish the differences across cultures, but also in various fields. For each real-world application a new dataset is necessary in order to gain information about the perspective of these different angles.

7.5.2 Usage of Multi-model Analysis

Another method can be the use of multi-model analysis. Facial glow and beauty is a very complex phenomenon, which can be influenced by different indicators, such as health, culture and social believes. These indicators can be interrelated and difficult to distangle using traditional facial image classification techniques. Further research is needed in the field of multi-model analysis methods to investigate the performance and the interpretability of the classification of both facial glow and beauty.

One method is utilising a multi-model analysis to combine different data sources. For instance, combining facial images from social media platforms with physiological data obtains a more comprehensive understanding of the concepts. But also the use of cultural or social information can improve the interpretability of the results and the facial glow or beauty analysis, such as including the age and gender of the perceiver to cluster different images in order to make a more accurate prediction.

7.5.3 Cross-cultural Analysis

Facial beauty and glow are subjective and highly influenced by cultures and regions. A cross-cultural analysis can provide insights in the perception of different cultures for these concepts. One method to account for this is comparing the facial beauty and glow perceptions across cultures and identify the differences in preferences. This analysis can be used to gain a more comprehensive understanding of what causes these preferences and perceptions.

Western beauty ideals include symmetry and the golden ratio, which can not be applied to facial images of people of other cultures. Moreover, for facial glow pigmentation signs can be an indicator of a decrease in radiance for some Asian countries, while this might not be the case for Western or

African countries. The differences, which can be derived from the cross-cultural analysis, can be used for highly effective marketing campaigns and more personalised skincare products. The one-fit-all approach does not fit in the current society given the high globalisation and diversity.

Chapter 8

Conclusions

8.1 Introduction

This thesis project aimed to answer the following research question:

> How can classification algorithms be developed for facial beauty and glow while incorporating cross-cultural complexity?

The aim of this study was to produce algorithms to automatically classify facial beauty and glow, including the cross-cultural complexity. The subs-questions are as follows:

- **Sub-question 1.** What are the criteria for the classification of facial beauty and facial glow?
- **Sub-question 2.** How is the social-cultural context influencing the requirements of the perceptions of facial beauty and glow?
- **Sub-question 3.** What are existing computational approaches for detecting facial beauty and glow (and how are they influenced by these social-cultural bias)?
- **Sub-question 4.** What are ethical implications of the classification of facial beauty and glow?

In this chapter the conclusion based on the different experiments are discussed and in parallel the questions as mentioned above are answered.

8.2 The Objective Criteria for facial beauty and glow

Facial beauty and glow are both very subjective notions, which have varied heavily over time. The technological revolution has brought many insights on these perception, instead of the traditional theories on these concepts. The emergence of social media and medical treatments has made beauty more inclusive and diverse, making it possible to create a cross-cultural perception of beauty and glow.

Based on the literature the following criteria are set for the definition of facial beauty:

1. Golden Ratio: the mathematical ratio between two features.
2. Facial Symmetry: the symmetry level of the face.
3. Skin Quality: absence of acne and other irregularities of the skin.

4. Facial Features: the form of facial features.
5. Facial Averageness: the averageness level of the face.

And similar criteria are set for facial glow, namely:

1. Luminosity of the skin: the overall radiance of the skin, especially in the T-zone.
2. Smoothness of the skin: the absence of wrinkles and imperfections of the skin.
3. Hydration of the skin: the hydrated appearance of the skin.
4. Evenness in skin-colour: the evenness in colour without any hyperpigmentation of the skin.
5. Elasticity of the skin: the ability of the skin to stretch and return.
6. Visibility of the pores: the absence of blackheads and facial pores.

Both notions seem to have criteria in common, such as the smoothness of the skin. However, facial glow is highly associated with health, while facial beauty does not necessary correlate with health. Skin quality and appearance is more important for the perception of facial glow, while other attributes such as bigger lips, averageness and right proportions are important for the perception of facial beauty.

The requirements given above are highly influenced by social-cultural context. For instance, the golden ratio has a Western bias, due to the fact that it is inapplicable to Malaysian faces for example. Additionally, in the annotation experiments, Chapter 6, we have seen that annotators from different age ranges do not reach an agreement level based on the calculated Fleiss' Kappa. However, when looking at the different age segments we do see a higher Fleiss' Kappa coefficients, which indicates a higher agreement between the annotates. While the number of participants was very low, 29 participants in total, and similarly for the number of annotated images (50 images in total), it is possible to assume that the age of the participants is an influence in their perception of facial glow. Further research is necessary in order to investigate the correlation between age and ethnicity and the perception of facial glow and beauty.

8.3 Deep Learning Approach Facial Beauty and Glow

8.3.1 The Highly Subjective Notion of Facial Beauty

In order to classify facial beauty images automatically, 8 deep learning models are trained with different hyperparameter settings. The models are trained in a traditional way, but also with the use of a Siamese network, where the similarity between the images are calculated, making it able for the model to pairwise compare the images.

Based on the Neural Network experiments, it can be concluded that the VGG19 model with RMSprop is the best model for facial beauty classification. The model is performing as one of the best on the different performance metrics. Moreover, the model performance is also stable on different batches compared to the other models. The other neural networks trained in the experiment showed a high risk of overfitting, while the extra dense layers prevented the VGG19 model of overfitting too much. Hence, this model has a better ability to generalise compared to the other model

The performance of the VGG19 model reaches a score of approximately 75% for recall, precision, accuracy and f1-score for each of the labels. For each 100 samples, the model is wrongly classifying 25 labels. The performance of the model does not increase and after a while the loss function starts to increase. The data used for the training are the female images of the CelebA dataset, which contains a lot of noise. This can be a cause of the slightly increasing loss function. Moreover, the performance of the model can also be explained due to the complexity of this particular task. The concept of facial beauty is a very subjective concept. This concept varies highly over time, ages, cultures and regions. The CelebA dataset is annotated by multiple annotators and the label *Attractive* had the lowest agreement score. This means that the ground truth is rather subjective and can vary, making it complex for the model to convergence or to learn the right patterns in the dataset.

For the second part of the experiment Siamese networks are used in order to predict the similarity between the images. The Siamese networks performed significantly worse than the neural networks. This can be explained due to the fact that the classification of facial beauty is highly subjective, and that facial beauty should be determined based on one input image. Comparing images on facial beauty remains very difficult and even humans can not perform this task correctly. If there is such thing as classifying it correctly. The images are not classified on whether someone has black hair or white hair, but on a highly subjective notion, namely beauty. Making it very complex for the Siamese network to calculate the similarity.

8.3.2 The Semi-Objective Notion of Facial Glow

Facial glow is a new field in computer vision, because no previous research has been conducted on this topic. This also means that no previous dataset was available to train the models. In order to be able to automatically classify facial glow images different steps have been taken. First of all, the notion of facial glow has been studied. Facial glow is highly correlated with a healthy appearance, including indicators such as hydration of the skin, evenness of the skin colour and the elasticity of the skin. The indicators are used to annotate glowing and non-glowing images based on the images used for the facial beauty classification.

A small subset with annotated images are used, namely a total number of 2000 images. These images were solely female images, with the same illumination. The same models (8 models in total) are trained on these images, both with the traditional method and the Siamese networks. The best performing model remains the VGG19 model, performing the best on all performance metrics, such as the accuracy, precision, f1-score and accuracy. Moreover, the VGG19 model resulted in an accuracy of 85%, which is higher compared to the facial beauty classification. The reason for the difference in the performance of the two classification task can be explained due to the number of annotators and a more solid ground truth for the facial glow annotations. A risk that occurs in this case is however the bias, due to the fact that one single annotators has labelled the images. Using more annotators can decrease the performance of the model due to the subjectivity of the concept.

However, even though facial glow is subjective, it may be more objective compared to the concept of facial beauty. Facial beauty is influenced by highly individual preferences, such as sexual preferences, cultural and social biases, but also age biases. While facial glow, due to its association with a healthy appearance, tend to be slightly more objective. Future research is necessary to gain deeper understanding of features that indicate facial glow.

Similar to the facial beauty classification. The second part of the experiment used the Siamese networks in order to determine the similarity between the images. The model results for the Siamese models for facial glow are comparable to the results for facial beauty. Based on this, it can be concluded that facial glow is also a highly complex task, making it impossible for the model to correctly compare the images and predict the right label. Therefore, Siamese Networks are not a suitable model for the task at hand.

8.4 Annotation Experiments

Chapter 6 presents the results of the annotation experiment. This experiment was a side project with the goal to understand the most important requirements for the perception of facial glow. Before executing the experiment, a correlation analysis was performed on the images for the attribute facial glow in order to identify patterns. The highest correlated features with the attribute

Attractive were *Heavy Makeup* and *Wearing Lipstick*. The lowest correlated attributes were all correlated with facial appearances. The dataset is highly biased towards female attractiveness. For this experiment only female faces were selected for the annotation task.

The annotation results are based on the labelling of participants with different backgrounds and ages. The majority of the participants were Dutch and young adults (25 - 34 years). The results showed that facial glow had no highly correlated features and no agreement level between the participants, which indicates a very subjective concept and heavily dependent on individual preferences.

However, when the participants were divided in different subsets within their own age category a certain agreement level was reached, meaning that the perception of facial glow is age dependent. Moreover, when asking the participants whether they associated facial glow with facial attractiveness, the answer was negative. Facial glow was not associated with facial beauty by the majority of the participants.

To conclude, the perception of facial glow is highly subjective based on the Fleiss' kappa coefficient of the annotations. Comparing the annotations of the total participant group no agreement level was reached, while at the level of smaller subset there was an agreement (similar age groups). However, these results are not representative due to the small sample group both in participants and the number of images. In order to gain a more comprehensive understanding of the perception of facial glow a larger experiment is necessary. Images from social media can be used in order to create a more realistic experiment. Moreover, the diversity of the participants need to also be increased in both age, gender and ethnicity. Lastly, this experiment was done based on female images. However, the perception of facial glow between female images and male images can differ significantly. Therefore, a more thorough experiment needs to be executed in order to gain deeper insights.

8.5 Native Language Search

The experiment of the native language search researched the significance of studying the perception of facial beauty and glow across various countries. The methodology for this experiment is the usage of a VPN to access the available servers in NordVPN. A total of 30 countries were selected to conduct the native language search. The native language search includes not only the respective language of the country, but also the most popular search engine. The results give insights in the disparities and commonalities in beauty standards and skincare practices between 6 regions in the world.

The implications of this experiment are both important for both individuals and the beauty industry as a whole. One main conclusion drawn based on the results is the necessity of a more inclusive and diverse approach to beauty, which includes and acknowledges different beauty ideals and skincare practices of different regions in the world. As any other experiment, this research also has its own limitations. One of them is to provide a more comprehensive understanding of facial beauty or glow by researching the differences between gender or expanding the countries.

This experiment highlights the importance of investigating beauty standards and practices across diverse cultures and regions to foster a more inclusive and diverse approach to beauty, and to challenge the narrow and often unrealistic beauty ideals that dominate mainstream media and advertising. As a concluding remark, we suggest that future research in this domain should prioritise the various perceptions of people from varied backgrounds and cultures to gain a more comprehensive understanding of the beauty standards and practices that shape our world, and to work towards a more inclusive and diverse vision of beauty.

8.6 The Ethical Implications of Classification

Artificial intelligence can reduce the amount of manual work, but also help humanity to gather deeper insights in notions such as facial beauty and glow. The use of new algorithms for either marketing campaigns or the production of new products, should however be applied with great care.

First of all, the algorithms are trained on a dataset, which are labelled by human annotators. These human annotators can be biased, either influenced by their age or culture, and therefore the annotation also will be biased. Hence, the algorithm trained on the biased dataset will become itself bias. The bias can lead to discrimination which only strengthens the inequality in our society. Moreover, the bias can also lead to an ideal image of facial beauty or glow, which might differ from the perception of smaller social groups. This can lead to a decrease in self esteem and other negative consequences for individuals. Additionally, to classify someone as non-attractive can be very unethical, because an ideal image is forced on this individual, which may not be applicable or true. We have to use the algorithms with great care and take in consideration that the distinction between attractive and non-attractive may not maintainable and even unethical.

Secondly, as seen in the results the networks do not perform optimally. The performance of the models can be caused by different factors, such as the model architecture, the underlying data, the amount of training time, but also on the subjective nature of the concepts facial beauty and glow. The question here is: Should we actually desire to objectify facial beauty or glow? Facial glow should be limited to objective patterns such as the associations with health, while facial beauty is a very personal and individualistic desire. The beauty can also lie in the eye of the beholder.

Bibliography

[Abolhasani et al., 2021] Abolhasani, R., Araghi, F., Tabary, M., Aryannejad, A., Mashinchi, B., and Robati, R. M. (2021). The impact of air pollution on skin and related disorders: a comprehensive review. *Dermatologic Therapy*, 34(2):e14840. 12

[Ahadit and Jatoth, 2022] Ahadit, A. B. and Jatoth, R. K. (2022). A novel multi-feature fusion deep neural network using hog and vgg-face for facial expression classification. *Machine Vision and Applications*, 33(4):1–23. 29

[Alam et al., 2015] Alam, M. K., Mohd Noor, N. F., Basri, R., Yew, T. F., and Wen, T. H. (2015). Multiracial facial golden ratio and evaluation of facial appearance. *PloS one*, 10(11):e0142914. 8

[Alippi et al., 2018] Alippi, C., Disabato, S., and Roveri, M. (2018). Moving convolutional neural networks to embedded systems: The alexnet and vgg-16 case. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 212–223. 29

[Alley and Cunningham, 1991] Alley, T. R. and Cunningham, M. R. (1991). Article commentary: averaged faces are attractive, but very attractive faces are not average. *Psychological science*, 2(2):123–125. 11

[Alom et al., 2018] Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S., and Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. 29

[Andini et al., 2022] Andini, C., Sosrohadi, S., Fairuz, F., Dalyan, M., Rahman, F. F., and Hasnia, H. (2022). The study of japanese women in the facial treatment advertisement: A semiotics perspective of pierce's theory. *ELS Journal on Interdisciplinary Studies in Humanities*, 5(2):337–347. 12

[Arpit et al., 2017] Arpit, D., Jastrzcebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. (2017). A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR. 29

[Aruleba and Viriri, 2021] Aruleba, I. and Viriri, S. (2021). Enhanced convolutional neural network for age estimation. In *Advances in Computational Intelligence: 16th International Work-Conference on Artificial Neural Networks, IWANN 2021, Virtual Event, June 16–18, 2021, Proceedings, Part I 16*, pages 385–394. Springer. 23

[Barrachina, 2022] Barrachina, J. (2022). Lightweight models in face attribute recognition: Performance under oclussions. In *17th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2022): Salamanca, Spain, September 5–7, 2022, Proceedings*, volume 531, page 270. Springer Nature. 23

[Bengio, 2012] Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In Guyon, I., Dror, G., Lemaire, V., Taylor, G., and Silver, D., editors, *Proceedings of*

ICML Workshop on Unsupervised and Transfer Learning, volume 27 of *Proceedings of Machine Learning Research*, pages 17–36. PMLR. 27

[Bengio et al., 1994] Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166. 29

[Blondell, 2013] Blondell, R. (2013). *Helen of Troy: beauty, myth, devastation.* Oxford University Press. 5

[Bougourzi et al., 2022] Bougourzi, F., Dornaika, F., and Taleb-Ahmed, A. (2022). Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. *Knowledge-Based Systems*, 242:108246. 17

[Broer et al., 2012] Broer, P. N., Buonocore, S., Morillas, A., Liu, J., Tanna, N., Walker, M., Ng, R., and Persing, J. A. (2012). Nasal aesthetics: a cross-cultural analysis. *Plastic and reconstructive surgery*, 130(6):843e–850e. 10

[Cao et al., 2020] Cao, K., Choi, K.-n., Jung, H., and Duan, L. (2020). Deep learning for facial beauty prediction. *Information*, 11(8):391. 17, 29

[Cao et al., 2017] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2017). Vggface2: A dataset for recognising faces across pose and age. 24

[Chen et al., 2021] Chen, Y., Zhang, Y., Huang, Z., Luo, Z., and Chen, J. (2021). Celebhair: A new large-scale dataset for hairstyle recommendation based on celeba. In *Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Tokyo, Japan, August 14–16, 2021, Proceedings, Part III*, pages 323–336. Springer. 23

[Chollet, 2017] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258. 31

[Cohen Jr, 1991] Cohen Jr, M. (1991). Picasso and the art of distortion and dislocation: the artist as researcher and experimentalist. *Journal of Craniofacial Genetics and Developmental Biology*, 11(2):61–95. 6

[Cunningham, 1986] Cunningham, M. R. (1986). Measuring the physical in physical attractiveness: Quasi-experiments on the sociobiology of female facial beauty. *Journal of personality and social psychology*, 50(5):925. 10

[Dibeklioğlu et al., 2014] Dibeklioğlu, H., Gevers, T., Lucassen, M., and Salah, A. A. (2014). The role of color and contrast in facial age estimation. In *International Workshop on Human Behavior Understanding*, pages 52–61. Springer. 18

[Dimitrov et al., 2023] Dimitrov, D., Maymone, M. B., and Kroumpouzos, G. (2023). Beauty perception: A historic and contemporary review. *Clinics in Dermatology*. 6, 7

[Dobai et al., 2018] Dobai, A., Markella, Z., Vízkelety, T., Fouquet, C., Rosta, A., and Barabás, J. (2018). Landmark-based midsagittal plane analysis in patients with facial symmetry and asymmetry based on cbct analysis tomography. *Journal of Orofacial Orthopedics/Fortschritte der Kieferorthopadie*, 79(6). 9

[Dobke et al., 2006] Dobke, M., Chung, C., and Takabe, K. (2006). Facial aesthetic preferences among asian women: are all oriental asians the same? *Aesthetic Plastic Surgery*, 30:342–347. 52

[Donahue et al., 2013] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2013). Decaf: A deep convolutional activation feature for generic visual recognition. 27

[Du et al., 2022] Du, Y., Doraiswamy, C., Mao, J., Zhang, Q., Liang, Y., Du, Z., Vasantharaghavan, R., and Joshi, M. K. (2022). Facial skin characteristics and concerns in indonesia: A cross-sectional observational study. *Skin Research and Technology*, 28(5):719–728. 16

[Dunlap, 1997] Dunlap, R. A. (1997). *The golden ratio and Fibonacci numbers.* World Scientific. 8

[Eggerstedt et al., 2020] Eggerstedt, M., Rhee, J., Urban, M. J., Mangahas, A., Smith, R. M., and Revenaugh, P. C. (2020). Beauty is in the eye of the follower: facial aesthetics in the age of social media. *American Journal of Otolaryngology*, 41(6):102643. 7

[Eisenthal et al., 2006] Eisenthal, Y., Dror, G., and Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. *Neural computation*, 18(1):119–142. 9, 18

[Ekmekji, 2016] Ekmekji, A. (2016). Convolutional neural networks for age and gender classification. *Stanford University.* 18

[El-Kilany and Raoof, 2017] El-Kilany, E. and Raoof, E. (2017). Facial cosmetics in ancient egypt. *Egyptian Journals of El-Kilany, Engy, Dan Enas Raoof.* 12

[Eldridge, 2015] Eldridge, L. (2015). *Face paint: The story of makeup.* Abrams. 6

[Erdman and Poutahidis, 2014] Erdman, S. and Poutahidis, T. (2014). Probiotic 'glow of health': it's more than skin deep. *Beneficial microbes*, 5(2):109–119. 12

[Fan et al., 2012] Fan, J., Chau, K., Wan, X., Zhai, L., and Lau, E. (2012). Prediction of facial attractiveness from facial proportions. *Pattern Recognition*, 45(6):2326–2334. 10

[Farahani et al., 2020] Farahani, A., Pourshojae, B., Rasheed, K., and Arabnia, H. R. (2020). A concise review of transfer learning. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 344–351. IEEE. 28

[Fernandez, 2001] Fernandez, L. B. (2001). *An exporation and analysis of the sociological content in the art of Frida Kahlo: Mexican ethnic and cultural identity, feminism, the struggle of the disable and her political ideology.* PhD thesis, California State University, Dominguez Hills. 6

[Fink et al., 2001] Fink, B., Grammer, K., and Thornhill, R. (2001). Human (homo sapiens) facial attractiveness in relation to skin texture and color. *Journal of Comparative Psychology*, 115(1):92. 10

[Fink and Neave, 2005] Fink, B. and Neave, N. (2005). The biology of facial beauty. *International journal of cosmetic science*, 27(6):317–325. 9, 10, 11, 16

[Fink et al., 2006] Fink, B., Neave, N., Manning, J. T., and Grammer, K. (2006). Facial symmetry and judgements of attractiveness, health and personality. *Personality and Individual differences*, 41(3):491–499. 9

[Flament et al., 2020] Flament, F., Abric, A., Amar, D., Ye, C., Caron, J., and Negre, C. (2020). Changes in facial signs due to age and their respective weights on the perception of age, on a tired-look or a healthy glow among differently aged Chinese men. *International Journal of Cosmetic Science*, 42(5):452–461. 2

[Flament et al., 2013] Flament, F., Bazin, R., Laquieze, S., Rubert, V., Simonpietri, E., and Piot, B. (2013). Effect of the sun on visible clinical signs of aging in caucasian skin. *Clinical, cosmetic and investigational dermatology*, pages 221–232. 10, 17

[Flament et al., 2021] Flament, F., Ye, C., Mercurio, D. G., Abric, A., Sewraj, P., Velleman, D., Yamamoto, S., Prunel, A., and Colomb, L. (2021). Evaluating the respective weights of some facial signs on the perceived radiance/glow in differently aged women of six countries. *Skin Research and Technology*, 27(6):1116–1127. 15, 17

[Gan et al., 2022] Gan, J., Xie, X., Zhai, Y., He, G., Mai, C., and Luo, H. (2022). Facial beauty prediction fusing transfer learning and broad learning system. *Soft Computing*, pages 1–14. 30, 31

[Genovese et al., 2017] Genovese, L., Corbo, A., and Sibilla, S. (2017). An insight into the changes in skin texture and properties following dietary intervention with a nutricosmeceutical containing a blend of collagen bioactive peptides and antioxidants. *Skin pharmacology and physiology*, 30(3):146–158. 10

[Goldie et al., 2021] Goldie, K., Kerscher, M., Fabi, S. G., Hirano, C., Landau, M., Lim, T. S., Woolery-Lloyd, H., Mariwalla, K., Park, J.-Y., and Yutskovskaya, Y. (2021). Skin quality–a holistic 360 view: consensus results. *Clinical, cosmetic and investigational dermatology*, pages 643–654. 17

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press. 29

[Grammer and Thornhill, 1994] Grammer, K. and Thornhill, R. (1994). Human (homo sapiens) facial attractiveness and sexual selection: the role of symmetry and averageness. *Journal of comparative psychology*, 108(3):233. 9, 11, 18

[Groathouse et al., 2010] Groathouse, J., Barcena, H., Gustafson, C., Wilson, M. V., Symonsbergen, D. J., Lucas, K. A., and Holmes, A. E. (2010). Wild plum: Novel particles of improved optical brightness and fluorescence. *International Journal of Cosmetic Science*, 32(6):472–473. 14

[Guinot et al., 2001] Guinot, C., Latreille, J., Tenenhaus, M., and Malvy, D.-M. (2001). Global classification of human facial healthy skin using pls discriminant analysis and clustering analysis. *International journal of cosmetic science*, 23(2):67–73. 16

[Gunes, 2011] Gunes, H. (2011). A survey of perception and computation of human beauty. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 19–24. 17

[Gunes and Piccardi, 2006a] Gunes, H. and Piccardi, M. (2006a). Assessing facial beauty through proportion analysis by image processing and supervised learning. *International Journal of Human-Computer Studies*, 64(12):1184–1199. 8

[Gunes and Piccardi, 2006b] Gunes, H. and Piccardi, M. (2006b). Assessing facial beauty through proportion analysis by image processing and supervised learning. *International Journal of Human-Computer Studies*, 64:1184–1199. 18

[Gunning and Aha, 2019] Gunning, D. and Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58. 28

[Haeri et al., 2019] Haeri, M., Phamduy, T., Cafone, N., Turkileri, K., and Velkov, D. (2019). Novel digital image analysis using fractal dimension for assessment of skin radiance. *Skin Research and Technology*, 25(4):564–571. 16

[Halberstadt and Rhodes, 2000] Halberstadt, J. and Rhodes, G. (2000). The attractiveness of nonface averages: Implications for an evolutionary explanation of the attractiveness of average faces. *Psychological Science*, 11(4):285–289. 11

[Halberstadt and Rhodes, 2003] Halberstadt, J. and Rhodes, G. (2003). It's not just average faces that are attractive: Computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin & Review*, 10(1):149–156. 11

[Harguess and Aggarwal, 2011] Harguess, J. and Aggarwal, J. K. (2011). Is there a connection between face symmetry and face recognition? In *CVPR 2011 WORKSHOPS*, pages 66–73. IEEE. 9

[Harrison and Hefner, 2014] Harrison, K. and Hefner, V. (2014). Virtually perfect: Image retouching and adolescent body image. *Media Psychology*, 17(2):134–153. 13

[Hassaballah et al., 2013] Hassaballah, M., Murakami, K., and Ido, S. (2013). Face detection evaluation: a new approach based on the golden ratio ϕ. *Signal Image Video Process.*, 7(2):307–316. 8

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. 30

[He et al., 2016a] He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. 27

[He et al., 2016b] He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 29, 30

[Heidari and Fouladi-Ghaleh, 2020] Heidari, M. and Fouladi-Ghaleh, K. (2020). Using siamese networks with transfer learning for face recognition on small-samples datasets. In *2020 International Conference on Machine Vision and Image Processing (MVIP)*, pages 1–4. IEEE. 32

[Henriques and Patnaik, 2020] Henriques, M. and Patnaik, D. (2020). Social media and its effects on beauty. In *Beauty-Cosmetic Science, Cultural Issues and Creative Developments*. IntechOpen. 7

[Heppt and Vent, 2015] Heppt, W. J. and Vent, J. (2015). The facial profile in the context of facial aesthetics. *Facial Plastic Surgery*, 31(05):421–430. 6

[Huang et al., 2013] Huang, C., Liu, X., and Chen, Y. (2013). Facial asymmetry index in normal young adults. *Orthodontics & craniofacial research*, 16(2):97–104. 9

[Huang et al., 2017] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708. 31

[Huang et al., 2008a] Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008a). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*. 22, 23

[Huang et al., 2008b] Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008b). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*. 24

[Humphrey et al., 2021] Humphrey, S., Brown, S. M., Cross, S. J., and Mehta, R. (2021). Defining skin quality: clinical relevance, terminology, and assessment. *Dermatologic Surgery*, 47(7):974. 10

[Ikeda et al., 2021] Ikeda, H., Saheki, Y., Sakano, Y., Wada, A., Ando, H., and Tagai, K. (2021). Facial radiance influences facial attractiveness and affective impressions of faces. *International Journal of Cosmetic Science*, 43(2):144–157. 15, 16, 17

[J Iyer et al., 2021] J Iyer, T., Nersisson, R., Zhuang, Z., Joseph Raj, A. N., Refayee, I., et al. (2021). Machine learning-based facial beauty prediction and analysis of frontal facial images using facial landmarks and traditional image descriptors. *Computational Intelligence and Neuroscience*, 2021. 8

[Jang et al., 2017] Jang, K. S., Bayome, M., Park, J. H., Park, K.-H., Moon, H.-B., and Kook, Y.-A. (2017). A three-dimensional photogrammetric analysis of the facial esthetics of the miss korea pageant contestants. *The korean journal of orthodontics*, 47(2):87–99. 8

[Johnston and Franklin, 1993] Johnston, V. S. and Franklin, M. (1993). Is beauty in the eye of the beholder? *Ethology and Sociobiology*, 14(3):183–199. 11

[Kagian et al., 2008] Kagian, A., Dror, G., Leyvand, T., Meilijson, I., Cohen-Or, D., and Ruppin, E. (2008). A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. *Vision Research*, 48:235–243. 11, 18

[Kant, 2000] Kant, I. (2000). *Critique of the power of judgment.* Cambridge University Press. 1

[Karam and Zhu, 2015] Karam, L. J. and Zhu, T. (2015). Quality labeled faces in the wild (qlfw): a database for studying face recognition in real-world environments. In *Human Vision and Electronic Imaging XX*, volume 9394, pages 87–96. SPIE. 23

[Kartal and Polat, 2022] Kartal, M. S. and Polat, Ö. (2022). Segmentation of skin lesions using u-net with efficientnetb7 backbone. In *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5. IEEE. 31

[Kashmar et al., 2019] Kashmar, M., Alsufyani, M. A., Ghalamkarpour, F., Chalouhi, M., Alomer, G., Ghannam, S., El Minawi, H., Saedi, B., Hunter, N., Alkobaisi, A., et al. (2019). Consensus opinions on facial beauty and implications for aesthetic treatment in middle eastern women. *Plastic and Reconstructive Surgery Global Open*, 7(4). 10

[Khanna and Datta Gupta, 2002] Khanna, N. and Datta Gupta, S. (2002). Rejuvenating facial massage–a bane or boon? *International journal of dermatology*, 41(7):407–410. 12

[Kim, 2021] Kim, J. S. (2021). Fine wrinkle treatment and hydration on the facial dermis using hydrotoxin mixture of microbotox and microhyaluronic acid. *Aesthetic Surgery Journal*, 41(6):NP538–NP549. 12, 16

[Kingma and Dhariwal, 2018] Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. 24

[Konstan, 2014] Konstan, D. (2014). *Beauty: the fortunes of an ancient Greek idea.* Oxford University Press. 5

[Koon, 2021] Koon, M. (2021). *State of the Art Rhinoplasty Techniques.* Springer. 7

[Koury and Epker, 1992] Koury, M. E. and Epker, B. N. (1992). Maxillofacial esthetics: Anthropometrics of the maxillofacial region. *Journal of Oral and Maxillofacial Surgery*, 50(8):806–820. 19

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. 27, 29

[Krizhevsky et al., 2017] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90. 29

[Kryger, 2023] Kryger, M. H. (2023). A spectacular image of sleep. *Sleep Health: Journal of the National Sleep Foundation.* 12

[Langlois et al., 1987] Langlois, J. H., Roggman, L. A., Casey, R. J., Ritter, J. M., Rieser-Danner, L. A., and Jenkins, V. Y. (1987). Infant preferences for attractive faces: Rudiments of a stereotype? *Developmental psychology*, 23(3):363. 11

[Larrabee, 1997] Larrabee, W. F. (1997). Facial beauty: myth or reality? *Archives of Otolaryngology–Head & Neck Surgery*, 123(6):571–572. 7

[Laughter et al., 2023] Laughter, M. R., Anderson, J. B., Maymone, M. B., and Kroumpouzos, G. (2023). Psychology of aesthetics: beauty, social media, and body dysmorphic disorder. *Clinics in Dermatology*. 7

[Learned-Miller et al., 2016] Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H., and Hua, G. (2016). Labeled faces in the wild: A survey. *Advances in face detection and facial image analysis*, pages 189–248. 23

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–44. 28, 29

[Lee et al., 2021] Lee, A. J., De La Mare, J. K., Moore, H. R., and Umeh, P. C. (2021). Preference for facial symmetry depends on study design. *Symmetry*, 13(9):1637. 9

[Lipton et al., 2014] Lipton, Z. C., Elkan, C., and Narayanaswamy, B. (2014). Thresholding classifiers to maximize f1 score. *Machine Learning and Knowledge Discovery in Databases*. 4

[Little et al., 2011] Little, A. C., Jones, B. C., and DeBruine, L. M. (2011). Facial attractiveness: evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1571):1638–1659. 9, 15

[Little et al., 2002] Little, A. C., Penton-Voak, I. S., Burt, D. M., and Perrett, D. I. (2002). Evolution and individual differences in the perception of attractiveness: How cyclic hormonal changes and self-perceived attractiveness influence female preferences for male faces. *Advances in social cognition Volume 1: Facial Attractiveness*. 18

[Liu et al., 2016] Liu, S., Fan, Y.-Y., Samal, A., and Guo, Z. (2016). Advances in computational facial attractiveness methods. *Multimedia Tools and Applications*, 75:16633–16663. 9

[Liu et al., 2019] Liu, X., Li, T., Peng, H., Chuoying Ouyang, I., Kim, T., and Wang, R. (2019). Understanding beauty via deep facial features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0. 10

[Liu et al., 2015a] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015a). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738. 22, 23

[Liu et al., 2015b] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015b). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738. 24

[Livio, 2002] Livio, M. (2002). The golden ratio and aesthetics. *Plus Magazine*, 22. 5

[Loughran, 1991] Loughran, J. (1991). Ayurvedic skin care: a glowing complexion comes from the inside out. *East West*, 21(4):32–35. 12

[MacCallum and Widdows, 2018] MacCallum, F. and Widdows, H. (2018). Altered images: understanding the influence of unrealistic images and beauty aspirations. *Health Care Analysis*, 26:235–245. 13

[MacDorman et al., 2009] MacDorman, K. F., Green, R. D., Ho, C.-C., and Koch, C. T. (2009). Too real for comfort? uncanny responses to computer generated faces. *Computers in human behavior*, 25(3):695–710. 9

[Maske et al., 2019] Maske, A. O., Pandhare, M., and Ashwin, D. W. (2019). Formulation and evaluation of herbal face pack for glowing skin. *International Journal of Advances in Pharmaceuticsm*, 8(01):5184–5189. 17

[Matsubara, 2012] Matsubara, A. (2012). Differences in the surface and subsurface reflection characteristics of facial skin by age group. *Skin Research and Technology*, 18(1):29–35. 16

[Melekhov et al., 2016] Melekhov, I., Kannala, J., and Rahtu, E. (2016). Siamese network features for image matching. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 378–383. IEEE. 31, 32

[Menon, 2019] Menon, A. V. (2019). Cultural gatekeeping in cosmetic surgery: Transnational beauty ideals in multicultural malaysia. *Poetics*, 75:101354. 10

[Michalsky, 2019] Michalsky, F. (2019). Fairness criteria for face recognition applications. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 527–528. 23

[Miller, 2022] Miller, C. (2022). *THE FENTY EFFECT: A CASE STUDY ON THE FUSION OF CELEBRITY, LUXURY, AND INCLUSIVITY*. PhD thesis, University of Oregon. 12

[Mohamad et al., 2014] Mohamad, M., Sabbri, A., Jafri, M. M., and Omar, A. (2014). Correlation between near infrared spectroscopy and electrical techniques in measuring skin moisture content. In *Journal of Physics: Conference Series*, volume 546, page 012021. IOP Publishing. 14

[Morris, 1967] Morris, D. (1967). *The Ape: A Zoologist's Study of the Human Animal*. Jonathan Cape. 10

[Mukherjee et al., 2011] Mukherjee, P. K., Maity, N., Nema, N. K., and Sarkar, B. K. (2011). Bioactive compounds from natural resources against skin aging. *Phytomedicine*, 19(1):64–73. 17

[Naidoo and Birch-Machin, 2017] Naidoo, K. and Birch-Machin, M. A. (2017). Oxidative stress and ageing: the influence of environmental pollution, sunlight and diet on skin. *Cosmetics*, 4(1):4. 10

[Naini et al., 2006] Naini, F. B., Moss, J. P., and Gill, D. S. (2006). The enigma of facial beauty: esthetics, proportions, deformity, and controversy. *American journal of orthodontics and dentofacial orthopedics*, 130(3):277–282. 10

[Navarro, 2020] Navarro, N. (2020). *The Use of Cosmetic Products in Relation to Latina Women's Self-Esteem*. PhD thesis, The Chicago School of Professional Psychology. 7

[Noor and Evans, 2003] Noor, F. and Evans, D. C. (2003). The effect of facial symmetry on perceptions of personality and attractiveness. *Journal of Research in Personality*, 37(4):339–347. 9

[Olah et al., 2017] Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. *Distill*, 2(11):e7. 28

[Olson, 2009] Olson, K. (2009). Cosmetics in roman antiquity: substance, remedy, poison. *Classical World*, pages 291–310. 6

[Oranges et al., 2016] Oranges, C. M., Schaefer, K. M., Gohritz, A., Haug, M., and Schaefer, D. J. (2016). The mirror effect on social media self-perceived beauty and its implications for cosmetic surgery. *Plastic and Reconstructive Surgery Global Open*, 4(11). 7

[Packiriswamy et al., 2012] Packiriswamy, V., Kumar, P., and Rao, M. (2012). Identification of facial shape by applying golden ratio to the facial measurements: an interracial study in malaysian population. *North American Journal of Medical Sciences*, 4(12):624. 8

[Padgett, 2015] Padgett, P. (2015). *The green beauty rules: The essential guide to toxic-free beauty, green glamour, and glowing skin*. Simon and Schuster. 12

[Pallett et al., 2010] Pallett, P. M., Link, S., and Lee, K. (2010). New "golden" ratios for facial beauty. *Vision Research*, 50:149–154. 10

[Palma et al., 2017] Palma, P., Khodaei, I., and Vasilenko, I. (2017). Emerging trends in the aesthetics of rhinoplasty: from local to global. *PMFA*. 7

[Pan and Yang, 2010] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359. 28

[Parkhi et al., 2015] Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In Xie, X., Jones, M. W., and Tam, G. K. L., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press. 24

[Perrett et al., 1994] Perrett, D., May, K., and Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. nature [internet]. 11

[Perrett et al., 1999] Perrett, D. I., Burt, D., Penton-Voak, I. S., Lee, K. J., Rowland, D. A., and Edwards, R. (1999). Symmetry and human facial attractiveness. *Evolution and Human Behavior*, 20(5):295–307. 9, 18

[Perseo, 2002] Perseo, G. (2002). The" beauty" of homo sapiens sapiens: standard canons, ethnical, geometrical and morphological facial biotypes. an explained collection of frontal north-europíde contemporary beauty facial canons-part iii. *Virtual Journal of Orthodontics [serial online]*. 19

[Phillips et al., 1997] Phillips, P., Moon, H., Rauss, P., and Rizvi, S. (1997). The feret evaluation methodology for face-recognition algorithms. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 137–143. 24

[Pirello, 2001] Pirello, C. (2001). *Glow: A Prescription for Radiant Health and Beauty*. Penguin. 12

[Posamentier and Lehmann, 2011] Posamentier, A. S. and Lehmann, I. (2011). *The glorious golden ratio*. Prometheus Books. 8

[Prokopakis et al., 2013] Prokopakis, E. P., Vlastos, I. M., Picavet, V. A., Nolst Trenite, G., Thomas, R., Cingi, C., and Hellings, P. W. (2013). The golden ratio in facial symmetry. *Rhinology*, 51(1):18–21. 5

[Ramanigopal, 2015] Ramanigopal, C. (2015). The impact of stress management among dental professionals. *INTERNATIONAL JOURNAL OF EMERGING MULTIDISCIPLINARY RESEARCH*, 5(7). 12

[Rhodes, 2006] Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annu. Rev. Psychol.*, 57:199–226. 9

[Rhodes et al., 1998] Rhodes, G., Proffitt, F., Grady, J., and Sumich, A. (1998). Facial symmetry and the perception of beauty. *Psychonomic Bulletin and Review*, 5:659–669. 9

[Rhodes et al., 2001a] Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., and Akamatsu, S. (2001a). Attractiveness of facial averageness and symmetry in non-western cultures: In search of biologically based standards of beauty. *Perception*, 30(5):611–625. 10

[Rhodes et al., 2001b] Rhodes, G., Zebrowitz, L. A., Clark, A., Kalick, S. M., Hightower, A., and McKay, R. (2001b). Do facial averageness and symmetry signal health? *Evolution and Human Behavior*, 22(1):31–46. 9

[Ricketts, 1982] Ricketts, R. M. (1982). Divine proportion in facial esthetics. *Clinics in Plastic Surgery*, 9(4):401–422. 19

[Romm, 1989] Romm, S. (1989). The changing face of beauty. *Aesthetic plastic surgery*, 13:91–98. 5

[Rubenstein et al., 2002] Rubenstein, A. J., Langlois, J. H., and Roggman, L. A. (2002). What makes a face attractive and why: The role of averageness in defining facial beauty. *Facial attractiveness: Evolutionary, cognitive, and social perspectives*. 11

[Saeed and Abdulazeez, 2021] Saeed, J. N. and Abdulazeez, A. M. (2021). Facial beauty prediction and analysis based on deep convolutional neural network: A review. *Journal Of Soft Computing And Data Mining*, 2:1–12. 17, 18

[Saeed et al., 2022] Saeed, J. N., Abdulazeez, A. M., and Ibrahim, D. A. (2022). Fiac-net: Facial image attractiveness classification based on light deep convolutional neural network. In *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)*, pages 1–6. IEEE. 18

[Sakano et al., 2021] Sakano, Y., Wada, A., Ikeda, H., Saheki, Y., Tagai, K., and Ando, H. (2021). Human brain activity reflecting facial attractiveness from skin reflection. *Scientific reports*, 11(1):1–13. 15, 17

[Samizadeh, 2022] Samizadeh, S. (2022). Facial beauty. *Non-Surgical Rejuvenation of Asian Faces*, pages 3–20. 6, 7

[Scott, 2016] Scott, D. A. (2016). A review of ancient egyptian pigments and cosmetics. *Studies in Conservation*, 61(4):185–202. 12

[Seok-Chan, 2013] Seok-Chan, E. (2013). Brilliant hallyu and plastic surgery. *jkms*, 28(11):1561–1562. 52

[Shafik and Elseesy, 2003] Shafik, A. and Elseesy, W. R. (2003). Medicine in ancient egypt. *Medicine Across Cultures: History and Practice of Medicine in Non-Western Cultures*, pages 27–47. 6

[Shah and Chew, 2018] Shah, S. and Chew, S.-K. (2018). Efficacy and safety of topical depigmenting agent in healthy human fair skin female volunteers: A single-arm study. *Journal of cosmetic dermatology*, 17(5):830–839. 14

[Shah et al., 2022] Shah, S., Manry, S., Makino, E. T., and Mehta, R. C. (2022). Clinical assessment of a circadian-based antioxidant system combined with a comprehensive brightening serum in diverse subjects with moderate-to-severe facial hyperpigmentation. *Journal of Cosmetic Dermatology*, 21(5):2082–2088. 16

[Sheldon et al., 2021] Sheldon, K. M., Goffredi, R., and Corcoran, M. (2021). The glow still shows: Effects of facial masking on perceptions of Duchenne versus social smiles. *Perception*, 50(8):720–727. PMID: 34162278. 2

[Simonyan et al., 2013] Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*. 28

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*. 27, 29

[Singh et al., 2019] Singh, P., Vijayan, R., and Mosahebi, A. (2019). The golden ratio and aesthetic surgery. *Aesthetic Surgery Journal*, 39(1):NP4–NP5. 5

[Sisti et al., 2021] Sisti, A., Aryan, N., and Sadeghi, P. (2021). What is beauty? *Aesthetic Plastic Surgery*, 45(5):2163–2176. 7

[Sokolova et al., 2006] Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In Sattar, A. and Kang, B.-h., editors, *AI 2006: Advances in Artificial Intelligence*, pages 1015–1021, Berlin, Heidelberg. Springer Berlin Heidelberg. 4

[Susilo et al., 2022] Susilo, D., Putranto, T. D., and Santos, R. R. T. (2022). The strategy of digital marketing of bening's clinic through instagram. *PRofesi Humas Jurnal Ilmiah Ilmu Hubungan Masyarakat*, 7(1):109–129. 13

[Sveikata et al., 2011] Sveikata, K., Balciuniene, I., Tutkuviene, J., et al. (2011). Factors influencing face aging. literature review. *Stomatologija*, 13(4):113–116. 12

[Szegedy et al., 2017] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31. 30

[Szegedy et al., 2015a] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015a). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. 27

[Szegedy et al., 2015b] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015b). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9. 30, 31

[Tan and Le, 2019] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR. 30, 31

[Tan and Le, 2021] Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR. 30, 31

[Terhörst et al., 2020] Terhörst, P., Fährmann, D., Kolf, J. N., Damer, N., Kirchbuchner, F., and Kuijper, A. (2020). Maad-face: A massively annotated attribute dataset for face images. *CoRR*, abs/2012.01030. 2, 22, 24

[Thareja et al., 2022] Thareja, R., Thareja, R., and Singh, S. (2022). Face mask detection system using resnet-50. *International Journal of Applied Engineering Research*, 17(6):588–592. 29

[Thomas and Dixon, 2016] Thomas, J. R. and Dixon, T. K. (2016). A global perspective of beauty in a multicultural world. *JAMA Facial Plastic Surgery*, 18(1):7–8. 5

[Thornhill and Gangestad, 1993] Thornhill, R. and Gangestad, S. W. (1993). Human facial beauty: Averageness, symmetry, and parasite resistance. *Human nature*, 4:237–269. 11

[Thornhill and Gangestad, 1999] Thornhill, R. and Gangestad, S. W. (1999). Facial attractiveness. *Trends in cognitive sciences*, 3(12):452–460. 9

[Tiggemann and Slater, 2017] Tiggemann, M. and Slater, A. (2017). Facebook and body image concern in adolescent girls: A prospective study. *International Journal of Eating Disorders*, 50(1):80–83. 13

[Torrey and Shavlik, 2010] Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global. 28

[Van den Berghe and Frost, 1986] Van den Berghe, P. L. and Frost, P. (1986). Skin color preference, sexual dimorphism and sexual selection: A case of gene culture co-evolution? *Ethnic and Racial Studies*, 9(1):87–113. 10

[Vats and Maurya, 2022] Vats, A. S. and Maurya, S. (2022). Formulation and evaluation of cosmetic herbal face pack for glowing skin. *International Journal of Advances in Pharmaceutics*. 12

[Verhoeven et al., 2016] Verhoeven, T., Xi, T., Schreurs, R., Bergé, S., and Maal, T. (2016). Quantification of facial asymmetry: a comparative study of landmark-based and surface-based registrations. *Journal of Cranio-Maxillofacial Surgery*, 44(9):1131–1136. 9

[Vrbančič and Podgorelec, 2020] Vrbančič, G. and Podgorelec, V. (2020). Transfer learning with adaptive fine-tuning. *IEEE Access*, 8:196197–196211. 28

[Wahab et al., 2022] Wahab, F., Nasim, R., Sultana, M. Z., and Momin, L. (2022). Maskne: A new entity during the covid-19 pandemic. *Acta dermatovenerologica Croatica.* 15

[Watson et al., 2016] Watson, B., Broadbent, J., Skouteris, H., and Fuller-Tyszkiewicz, M. (2016). A qualitative exploration of body image experiences of women progressing through pregnancy. *Women and Birth*, 29(1):72–79. 15

[Wei et al., 2022] Wei, W., Ho, E. S., McCay, K. D., Damaševičius, R., Maskeliūnas, R., and Esposito, A. (2022). Assessing facial symmetry and attractiveness using augmented reality. *Pattern Analysis and Applications*, 25(3):635–651. 9

[Westmore, 2001] Westmore, M. G. (2001). Camouflage and makeup preparations. *Clinics in dermatology*, 19(4):406–412. 6

[Wilner, 1931] Wilner, O. L. (1931). Roman beauty culture. *The Classical Journal*, 27(1):26–38. 6

[Winkler, 1990] Winkler, J. J. (1990). Phallos politikos: Representing the body politic in athens. *differences*, 2(1):29–45. 12

[Wu et al., 2022] Wu, H., Bezold, G., Günther, M., Boult, T., King, M. C., and Bowyer, K. W. (2022). Consistency and accuracy of celeba attribute values. *arXiv preprint arXiv:2210.07356.* 23

[Wu et al., 2020] Wu, X., Liu, R., Yang, H., and Chen, Z. (2020). An xception based convolutional neural network for scene image classification with transfer learning. In *2020 2nd international conference on information technology and computer application (ITCA)*, pages 262–267. IEEE. 31

[Xu et al., 2017] Xu, J., Jin, L., Liang, L., Feng, Z., Xie, D., and Mao, H. (2017). Facial attractiveness prediction using psychologically inspired convolutional neural network (pi-cnn). In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1657–1661. IEEE. 18

[Xu et al., 2018] Xu, L., Xiang, J., and Yuan, X. (2018). Transferring rich deep features for facial beauty prediction. *arXiv preprint arXiv:1803.07253.* 27

[Yadav and Yadav, 2015] Yadav, N. and Yadav, R. (2015). Preparation and evaluation of herbal face pack. *International Journal of Recent Scientific Research*, 6(5):4334–4337. 12

[Yalta et al., 2016] Yalta, K., Ozturk, S., and Yetkin, E. (2016). Golden ratio and the heart: A review of divine aesthetics. *International journal of cardiology*, 214:107–112. 8

[Yang et al., 2020] Yang, Q., Zhang, Y., Dai, W., and Pan, S. J. (2020). *Transfer learning.* Cambridge University Press. 28

[Yip et al., 2019] Yip, J., Ainsworth, S., and Hugh, M. T. (2019). Beyond whiteness: Perspectives on the rise of the pan-asian beauty ideal. *Race in the marketplace: Crossing critical boundaries*, pages 73–85. 10

[Yosinski et al., 2014] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems.* 27

[Youn, 2019] Youn, A. (2019). What is the ideal instagram filter? In *Aesthetic Surgery Journal Open Forum*, volume 1, page ojz019. Oxford University Press US. 13

[Yu et al., 2020] Yu, J., Li, M., Hao, X., and Xie, G. (2020). Deep fusion siamese network for automatic kinship verification. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 892–899. IEEE. 32

[Yu et al., 2017] Yu, Z., Wang, X., Su, B., and Zhang, Y. (2017). First evidence of the use of freshwater pearls as a cosmetic in ancient china: Analysis of white makeup powder from a northern song dynasty lv tomb (lantian, shaanxi province, china). *Archaeometry*, 59(4):762–774. 6

[Zaidel et al., 1995] Zaidel, D. W., Chen, A. C., and German, C. (1995). She is not a beauty even when she smiles: Possible evolutionary basis for a relationship between facial attractiveness and hemispheric specialization. *Neuropsychologia*, 33(5):649–655. 11

[Zebrowitz and Rhodes, 2002] Zebrowitz, L. A. and Rhodes, G. (2002). Nature let a hundred flowers bloom: the multiple ways and wherefores of attractiveness. *Facial attractiveness: Evolutionary, cognitive, and social perspectives.* 11, 18

[Zeppieri, 2022] Zeppieri, S. (2022). *Asian Foods for Longevity: Secret to Glowing Skin.* Susan Zeppieri. 12

[Zhai et al., 2020] Zhai, Y., Huang, Y., Xu, Y., Gan, J., Cao, H., Deng, W., Labati, R. D., Piuri, V., and Scotti, F. (2020). Asian female facial beauty prediction using deep neural networks via transfer learning and multi-channel feature fusion. *IEEE Access*, 8:56892–56907. 17

[Zhang et al., 2016a] Zhang, D., Chen, F., and Xu, Y. (2016a). Data-driven facial beauty analysis: Prediction, retrieval and manipulation. *Computer Models for Facial Beauty Analysis*, page 217–234. 18

[Zhang et al., 2016b] Zhang, D., Chen, F., Xu, Y., et al. (2016b). *Computer models for facial beauty analysis.* Springer. 17

[Zhang et al., 2011] Zhang, D., Zhao, Q., and Chen, F. (2011). Quantitative analysis of human facial beauty using geometric features. *Pattern Recognition*, 44(4):940–950. 18

[Zhang et al., 2020] Zhang, J., Li, S., Zhang, J.-Y., Du, F., Qi, Y., and Liu, X. (2020). A literature review of the research on the uncanny valley. In *Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments: 12th International Conference, CCD 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*, pages 255–268. Springer. 9

[Zhang et al., 2019] Zhang, K., Guo, Y., Wang, X., Yuan, J., and Ding, Q. (2019). Multiple feature reweight densenet for image classification. *IEEE Access*, 7:9872–9880. 31

[Zhang et al., 2017a] Zhang, L., Zhang, D., Sun, M.-M., and Chen, F.-M. (2017a). Facial beauty analysis based on geometric feature: Toward attractiveness assessment application. *Expert Systems with Applications*, 82:252–265. 5

[Zhang and Sabuncu, 2018] Zhang, Z. and Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31. 29

[Zhang et al., 2017b] Zhang, Z., Song, Y., and Qi, H. (2017b). Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818. 22

[Zheng et al., 2021] Zheng, D. X., Mulligan, K. M., and Scott, J. F. (2021). Tiktok and dermatology: an opportunity for public health engagement. *Journal of the American Academy of Dermatology*, 85(1):e25–e26. 13

[Zhou et al., 2016] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929. 29

[Zhu and Newsam, 2017] Zhu, Y. and Newsam, S. (2017). Densenet for dense flow. In *2017 IEEE international conference on image processing (ICIP)*, pages 790–794. IEEE. 31

[Zwahlen et al., 2022] Zwahlen, R. A., Tang, A. T., Leung, W. K., and Tan, S. K. (2022). Does 3-dimensional facial attractiveness relate to golden ratio, neoclassical canons,'ideal'ratios and 'ideal'angles? *Maxillofacial Plastic and Reconstructive Surgery*, 44(1):28. 8

Appendix A

Facial Glow Across the World: Keywords and Phrases

Chapter 6 explains the experiment performed to investigate the different beauty and skincare standards across cultures and regions. One of the components of this experiment is the native language search. Various keywords and phrases of facial beauty and glow are used in search engines to conduct insights. Table A.1 and A.2 contains the keywords and phrases used in the experiment.

Country	Glowing Face	Radiant Face	Beautiful Face	Beautiful Face
The Netherlands	glimmend gezicht	stralend gezicht	mooi gezicht	prachtig gezicht
France	visage éclatant	visage lumineux	beau visage	magnifique visage
Italy	volto luminoso	viso raggiante	bel viso	
Spain	cara que brilla intensamente	cara resplandeciente	hermoso rostro	cara hermosa
Sweden	glödande ansikte	strålande ansikte	vackert ansikte	härligt ansikte
Austria	strahlendes Gesicht	leuchtendes Gesicht	schönes Gesicht	wunderschönes gesicht
Portugal	rosto brilhante	cara brilhante	rosto bonito	lindo rosto
Greece	όμορφο πρόσωπο	ωραίο πρόσωπο	λαμπερό πρόσωπο	
Germany	strahlendes Gesicht	leuchtendes Gesicht	schönes Gesicht	wunderschönes gesicht
Singapore	Wajah berseri (亮丽的脸庞)		kecantikan wajah 面部美容	keindahan wajah 脸蛋漂亮
Japan	輝く顔	輝いている顔	美しい顔	かわいい
Hong Kong	焕发光彩	面色红润	美貌	姣好的面容
Taiwan	皮肤光彩	焕发光彩的脸	美丽的脸	神采奕奕的脸
Vietnam	khuôn mặt rạng rỡ	mặt sáng	gương mặt đẹp	khuôn mặt đẹp khuôn mặt xinh đẹp mặt đẹp
Thailand	หน้าสวย	ใบหน้าที่สวยงาม	ใบหน้าที่เร่าร้อน	ใบหน้าเปล่งประกาย
South Korea	아름다운 얼굴	미인	빛나는 얼굴	빛나
Malaysia	Wajah berseri muka yang berseri	muka bercahaya	kecantikan wajah cantik muka	keindahan wajah Muka cantik
South Africa	gloeiende gesig	stralende gesig	Mooi gesig	pragtige gesig

Figure A.1: Keywords and phrases in local languages

United Arab Emirates	وجه مشع	وجه مشرق	وجه جميل	وجه جميل
Isreal	פנים מזוהרים	פנים זוהרים	פנים יפים	פנים יפים
Turkey	Parlayan Yüz	Işıltılı Yüz	Güzel Yüz	Güzel Yüz
Australia	Glowing Face	Radiant Face	Beautiful Face	Beautiful Face
New Zealand	Glowing Face	Radiant Face	Beautiful Face	Beautiful Face
United States of America	Glowing Face	Radiant Face	Beautiful Face	Beautiful Face
Canada	Glowing Face	Radiant Face	Beautiful Face	Beautiful Face
Brazil	rosto brilhante	cara brilhante	rosto bonito	lindo rosto
Argentina	cara que brilla intensamente	cara resplandeciente	hermoso rostro	cara hermosa
Mexico	cara que brilla intensamente	cara resplandeciente	hermoso rostro	cara hermosa
Costa Rica	cara que brilla intensamente	cara resplandeciente	hermoso rostro	cara hermosa
Chile	cara que brilla intensamente	cara resplandeciente	hermoso rostro	cara hermosa

Figure A.2: Keywords and phrases in local languages