Facial Attractiveness Prediction Using a Single and Multi-Task Vision Transformer Framework

Mohammad Soroush Ghorbanimehr

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science (Computer Science and Software Engineering) at

Concordia University

Montréal, Québec, Canada

September 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mohammad Soroush Ghorbanimehr**

Entitled: **Facial Attractiveness Prediction Using a Single and Multi-Task Vision Transformer Framework**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (Computer Science and Software Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Sudhir Mudur

_____ External Examiner
Dr. Yang Wang

_____ Examiner
Dr. Sudhir Mudur

_____ Supervisor
Dr. Ching Yee Suen

Approved by _____
Dr. Joey Paquet, Chair
Department of Computer Science and Software Engineering

_____ 2025

_____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Facial Attractiveness Prediction Using a Single and Multi-Task Vision Transformer Framework

Mohammad Soroush Ghorbanimehr

Facial attractiveness prediction is a challenging and inherently subjective task in computer vision, with applications spanning social media, cosmetic technology, and aesthetic medicine. While convolutional neural networks (CNNs) have driven significant advances in this area, recent developments in transformer-based architectures, such as the Vision Transformer (ViT), offer new opportunities by capturing global feature relationships and long-range dependencies within images.

This thesis explores the use of Vision Transformers for predicting facial attractiveness on the SCUT-FBP5500 dataset, where beauty scores are computed from the average ratings of multiple human annotators. The task is formulated as a regression problem to predict continuous attractiveness scores. To enhance the learned feature representations, a multi-task learning framework is introduced, jointly performing gender and ethnicity classification alongside beauty prediction.

The methodology includes systematic image preprocessing, transfer learning with a ViT pretrained on large-scale facial recognition data, and fine-tuning for both primary and auxiliary tasks. Model performance is evaluated using PC, MAE, and RMSE for regression and classification accuracy for auxiliary tasks. Comparative experiments with CNN-based baselines demonstrate that transformer architectures capture more holistic and subtle aesthetic cues, resulting in improved prediction consistency.

Experimental results show that the proposed ViT-based approach achieves superior accuracy and robustness compared to conventional CNNs, even with limited training data. These findings highlight the potential of our Vision Transformers as an effective and data-efficient alternative for

facial aesthetic analysis. The thesis concludes by emphasizing the value of multi-task learning in enriching feature representations and encourages future research toward interpretable and scalable beauty prediction systems.

Acknowledgments

First and foremost, I am deeply grateful to my parents, Mojgan and Saeid, and my family for their unwavering love, support, and encouragement throughout my academic journey. Their belief in me has been the foundation of my resilience, perseverance, and motivation.

I wish to express my sincere appreciation to my supervisor, Dr. Ching Yee Suen, for his invaluable guidance, expertise, and mentorship. His insightful feedback and patient support have shaped both the direction and depth of this research.

I am profoundly thankful to Sepehr, Babak, and my family for their kindness and constant support, which have always reminded me of the importance of family bonds. I also wish to thank Fahimeh, my teammate, for her collaboration and encouragement throughout this project, and Nicola, our research manager, for his advice, technical support, and thoughtful guidance. Finally, my heartfelt gratitude goes to Paria for her constant encouragement, patience, and unwavering support during this journey.

I gratefully acknowledge the contributions of the open-source community and the creators of the Vision Transformer (ViT), along with all the researchers whose work forms the foundation of this thesis.

To everyone who has accompanied and supported me on this journey, I am sincerely grateful. Your contributions have been essential to my progress, shaping not only my academic work but also my personal growth.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past few years, facial beauty assessment has gained considerable attention and has become a rapidly growing area of research across different domains. The automated evaluation of facial aesthetics offers a captivating area of study with a wide range of possible applications. This chapter provides an overview of the evolution and key insights into facial beauty research while also highlighting its future applications.

1.1 Overview

Facial beauty has gained increasing attention in recent years, significantly influencing social standards and personal self-perception. This rising focus has fueled an expanding market for cosmetic procedures and products, both surgical and non-surgical. This societal shift underscores the importance of facial aesthetics, not only in personal identity but also in broader social dynamics. While facial beauty prediction (FBP) is relatively new to the field of computer science, it has long been a focus in psychology (X. Liu et al., 2019) , and today, it garners interest from various disciplines, including computer vision and medicine (Eisenthal, Dror, & Ruppin, 2006a; S. Liu, Fan, Guo, Samal, & Ali, 2017a; S. Liu, Fan, Samal, & Guo, 2016).

With the rapid advancements in Artificial Intelligence (AI) and Deep Learning (DL), there is growing interest in developing objective methods to analyze facial attractiveness. Facial beauty prediction (FBP) refers to the ability to automatically assign a beauty score to a facial image, reflecting

how attractive the face is perceived (Kao, He, & Huang, 2016; Liang et al., 2018; S. Liu et al., 2017a).

The FBP process generally involves three primary steps: acquiring a facial image dataset, extracting relevant features, and developing a predictive model. In most approaches, human raters evaluate a set of facial images, typically assigning numerical scores to indicate attractiveness. These ratings form the ground truth, which is used to train predictive models (Liang et al., 2018). Features such as facial landmarks or proportions are extracted from the images, and then the model is trained to predict beauty scores for new, unseen images.

Traditional facial beauty prediction methods often rely on hand-crafted features like geometric measurements. However, these methods suffer from limitations, such as the need for manual intervention in landmark localization, and their inability to fully capture the complexities of facial aesthetics. Additionally, data scarcity and class imbalance pose significant challenges, particularly when dealing with multi-class attractiveness ratings (Vahdati & Suen, 2020).

This thesis focuses on addressing these challenges by leveraging state-of-the-art deep learning models—specifically, the Vision Transformer (ViT). The ViT excels in capturing global, long-range dependencies due to its self-attention mechanism, making it suitable for understanding the overall structure and symmetry of the face.

Moreover, the application of multi-task learning in our model extends its functionality beyond beauty score prediction. Our model is also designed to classify gender (male or female) and ethnicity (Asian or Caucasian), providing a holistic approach to facial analysis. This multi-task learning approach enables the model to handle related tasks efficiently, improving both performance and generalization. By sharing common representations learned from facial images, our model can extract features useful for all tasks, thus enhancing the accuracy of each individual task (Zhang, Luo, Loy, & Tang, 2014).

1.2 Main Goal

The primary goal of this research is to develop an advanced and unbiased framework using transformer models (ViT) for facial beauty score prediction, gender classification, and ethnicity

classification. We aim to use transfer-learning approach that addresses challenges in data scarcity and model performance.

The thesis aims to demonstrate that accurate beauty score prediction is achievable even with limited annotated data by leveraging transfer learning and transformer-based architectures. To this end, several key challenges were addressed:

- **Data Imbalance:** The dataset for classification tasks exhibited substantial imbalance, with notable disparities in sample counts across attractiveness categories. Initially, facial beauty prediction was approached as a classification problem, and mitigation strategies such as class weighting and data augmentation were applied. However, these methods produced suboptimal results. Consequently, the task was reformulated as a regression problem for beauty score prediction, enabling a continuous and more nuanced representation of attractiveness.

- **Preprocessing:** Standard preprocessing steps included resizing all images to a fixed resolution, normalizing pixel values, and applying data augmentation techniques such as random cropping, horizontal flipping and a custom facial component symmetry alignment to improve generalization and reduce overfitting.

- **Model Fine-tuning:** Multiple pre-trained models were evaluated, with fine-tuning strategies specifically designed to improve performance in facial beauty analysis. These models were first trained on large-scale datasets such as ImageNet, followed by facial recognition datasets such as VGGFace2, and were further specialized for this task through targeted training on beauty-related facial images.

- **Training Configurations:** Extensive experimentation was conducted with different learning rates, batch sizes, and augmentation strategies to identify the most effective training setup for stable convergence and optimal performance.

- **Attention Map Analysis:** Self-attention maps generated by the transformer models were analyzed to identify the facial regions most influential in the prediction process, thereby enhancing model interpretability and providing insights into learned feature representations.

By addressing these challenges, we aim to surpass the performance of existing methods and set a new standard for facial beauty prediction.

1.3 Contributions

This thesis makes several key contributions:

- We present an innovative multi-task learning model that performs beauty score prediction, gender classification, and ethnicity classification simultaneously using a Vision Transformer (ViT) architecture.
- The proposed model addresses the issue of data scarcity by leveraging transfer learning techniques, significantly improving prediction accuracy on limited datasets.
- Our model uses attention maps to provide insights into which facial features are most influential in beauty prediction, helping interpret model outputs.
- We offer a comprehensive evaluation of transformer models in facial beauty prediction, demonstrating their superiority over traditional CNN-based approaches in regression tasks.

1.4 Structure

The thesis is structured as follows:

- **Chapter One:** Introduces the research topic, outlines its significance, and presents the main objectives and contributions of the thesis.
- **Chapter Two:** Provides a comprehensive literature review on facial beauty prediction, covering traditional approaches, deep learning methods, and transformer-based models.
- **Chapter Three:** Describes the datasets used in this study, including their characteristics and preprocessing procedures.
- **Chapter Four:** Details the proposed methodology, including model architectures, training strategies, and implementation details.

- **Chapter Five:** Presents the experimental results and evaluates the performance of the models using relevant metrics.
- **Chapter Six:** Discusses the findings, highlights limitations, and suggests directions for future research.

Chapter 2

Literature Review

In this chapter, we begin by exploring the fundamental principles behind facial beauty. Next, we present a review of the current methods used for predicting facial attractiveness. Finally, we examine the application of multi-task learning and multi-region strategies in facial analysis.

2.1 Facial Attractiveness Prediction: Overview

2.1.1 Studies on Aesthetic Appeal

The concept of ideal proportions, often referred to as *beauty canons*, has been advocated by those who believe that beauty can be objectively measured and quantified. This notion dates back to ancient times, where the human form was often analyzed in terms of symmetry and proportion Laurentini and Bottino (2014); Yi, Lei, and Li (2015). Faces that adhere to these established beauty standards are generally perceived as attractive.

Among these standards, the *golden ratio* has played a significant role, being widely applied in various research studies focused on facial beauty. This mathematical principle, often linked to harmony and balance, is used to assess the aesthetic appeal of facial features, and its influence extends beyond mere physical appearance, impacting perceptions of beauty across different cultures and time periods.

Furthermore, modern computational models and algorithms frequently integrate these ratios to enhance the accuracy of beauty prediction systems, making the golden ratio a critical component in

both historical and contemporary beauty analysis.

2.1.2 Ratios Golden

In ancient times, many artists, including sculptors and painters, believed that facial beauty adhered to a specific mathematical principle known as the Golden ratio. This ratio, approximately 1.618 to 1, has been regarded as an ideal measurement for facial attractiveness across centuries. It suggests that certain proportional relationships between facial features contribute to an aesthetically pleasing appearance. For example, as illustrated in Figure 1, the vertical distance from the top of the face to the nose is compared to the vertical distance between the nostrils and the tip of the chin, reflecting this timeless concept of beauty.

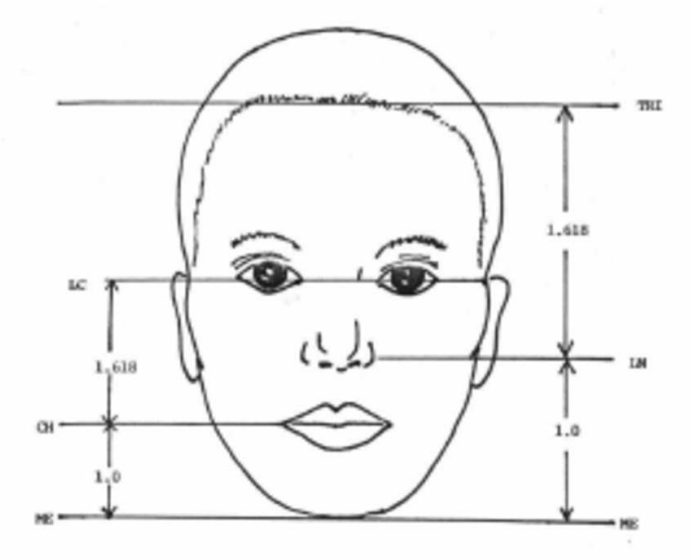

Figure 2.1: Reference image for the Golden ratio (sourced from (Bottino & Laurentini, 2010)).

An interesting example of the application of the golden ratio can be found in one of the most iconic pieces of art, Leonardo Da Vinci's Mona Lisa. This masterpiece is widely recognized not only for its artistic value but also for its mathematical precision, as it incorporates the principles of the golden ratio in its composition (see Figure 2). This subtle use of proportion contributes to the painting's harmonious and balanced aesthetic, making it a prime example of how art and mathematics intersect.

Furthermore, Figure 3 illustrates the Phi mask, a facial template designed by Marquardt that

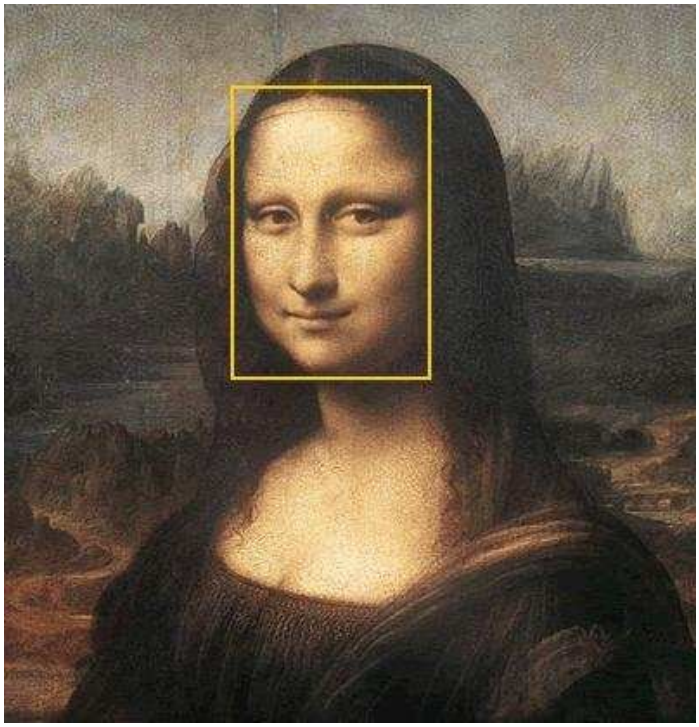

Figure 2.2: Depiction of the Golden ratio in the renowned painting Mona Lisa (sourced from (Saari et al., 2008)).

represents an idealized face structure based on the golden ratio. This template has been widely used in both scientific studies and aesthetic fields to analyze facial symmetry and attractiveness, as it is believed to reflect the proportions that are naturally perceived as beautiful. Marquardt's Phi mask serves as a visual guide to understanding how the golden ratio can be applied to the human face to achieve an optimal sense of balance and harmony

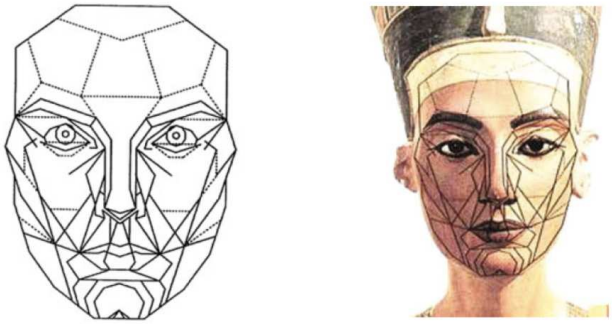

Figure 2.3: a: The Phi mask, developed by Marquardt, b: The Egyptian queen Neferneferuaten Nefertiti

2.1.3 Neoclassical Canons

Emerging during the Renaissance period, neoclassical canons were proposed by artists and scholars aiming to define the ideal proportions of the human face. These canons established a set of guidelines for facial aesthetics, grounded in the belief that beauty is closely linked to symmetry and mathematical ratios (Bozkir et al., 2004; J. Fan, Chau, Wan, Zhai, & Lau, 2012; Schmid, Marx, & Samal, 2008b). Renowned figures such as Leonardo da Vinci and Albrecht Dürer were instrumental in advancing these principles, applying them in their artworks to capture harmony and balance.

Table 2.1 outlines six key neoclassical canons, each describing specific proportional relationships between facial features. Notably, canons 4, 5, 7, and 8 focus on the relationship between two features, while canons 2 and 6 involve three distinct measurements. For example, Canon 4 asserts that the nose length should equal the ear length, reflecting a harmonious correspondence. Similarly, Canon 5 emphasizes the proportionality between the interocular distance and the nose width, reinforcing the role of symmetry in facial attractiveness.

While these ratios were initially used in Renaissance art, their influence extends to modern scientific research in facial analysis. The neoclassical canons have served as a foundation for contemporary studies of facial aesthetics, contributing to the development of methods used in cosmetic surgery, facial reconstruction, and algorithmic beauty prediction models. Although these ratios present an idealized framework, it is essential to acknowledge the subjective nature of beauty, which is shaped by cultural norms and personal preferences. Nonetheless, the neoclassical canons continue to provide valuable insights into the mathematical structure underlying perceptions of facial beauty.

Table 2.1: Six Neoclassical Canons (Extracted from (Bozkir et al., 2004), full reference is provided in the main paper)

No.	**Description**
2	Forehead height = nose length = lower face height
4	Nose length = ear length
5	Interocular distance = nose width
6	Interocular distance = right or left eye fissure width
7	Mouth width = 1.5 $\times$ nose width
8	Face width = 4 $\times$ nose width

2.1.4 Vertical Thirds and Horizontal Fifths

Marcus Vitruvius, a famous Roman architect, introduced the facial thirds rule, which divides the face into three equal horizontal sections: from the hairline to the eyebrows, the eyebrows to the base of the nose, and the base of the nose to the chin (see Figure 2.5). This concept, rooted in Vitruvius' study of human symmetry, has influenced fields such as art, architecture, and modern facial aesthetics.

Today, the facial thirds rule is still used in areas like orthodontics, cosmetic surgery, and facial analysis models. While an idealized framework, it serves as a foundation for assessing facial symmetry and balance, key elements in beauty perception.

Figure 2.4: Facial thirds concept presented by Marcus Vitruvius(Bottino & Laurentini, 2010)

Likewise, the width of an aesthetically pleasing face can be divided into five equal sections (refer to Figure 2.5).

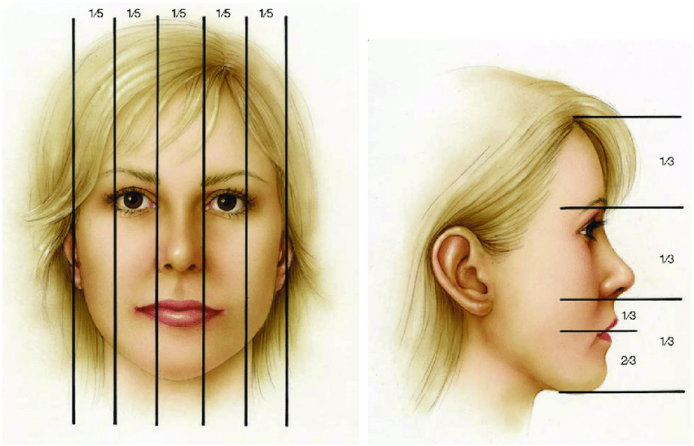

Figure 2.5: Illustration of the facial vertical thirds (left) and horizontal fifths (right) guidelines on a face)

2.1.5 Traditional learning model in facial beauty analysis

With the rapid advancements in pattern analysis and machine learning, computer-based methods for predicting facial attractiveness have gained increasing significance. These techniques have evolved to become powerful tools for automating beauty assessments, offering more objective and scalable approaches. Facial beauty prediction (FBP) methods can generally be categorized into two main types: traditional methods, which rely on handcrafted features and rules, and deep learning-based methods, which leverage neural networks to automatically extract features from facial images.

Despite the differences in methodology, most approaches in the literature on facial beauty assessment follow a similar overall framework (S. Liu et al., 2016), as depicted in Figure 8. This framework typically involves facial feature extraction, followed by model training and prediction, and is aimed at quantifying beauty in a consistent and reproducible manner.

In traditional facial beauty prediction (FBP) methods, the use of hand-crafted features alongside shallow predictors has been widespread. Among the various feature types, geometric features have been the most frequently applied in facial beauty research. These features are derived from the positions of facial landmarks and include measurements such as distances between key points on

the face, ratios of these distances, angles, and inclinations. Of these, distances between facial landmarks and the ratios formed by these measurements have become increasingly popular for their effectiveness in modeling attractiveness.

Table 2.2 provides a brief summary of the geometric features utilized in several existing studies on facial attractiveness. Identifying the most critical facial regions and their corresponding landmarks is often considered a fundamental step in many facial beauty analysis techniques (Laurentini & Bottino, 2014). By locating and analyzing these key points, researchers aim to capture the essential aspects of facial structure that contribute to perceived attractiveness.

For further details on facial attractiveness research and a comprehensive review of the topic, readers are encouraged to consult Refs. (Laurentini & Bottino, 2014) and (S. Liu, Fan, Guo, Samal, & Ali, 2017b), which provide valuable insights into recent advancements in the field.

Table 2.2: Summary of Extracted Geometric Features from Facial Landmarks

Reference	**Extracted Geometric Features**
Gunes and Piccardi Gunes and Piccardi (2006)	13 ratios of inter-landmark distances
Eisental et al. Eisenthal, Dror, and Ruppin (2006b)	37 distances and ratios from 36 facial landmarks
Kagian et al. Kagian et al. (2008)	6,972 geometric features (3,486 distances and 3,486 slopes), reduced to 90 using PCA, from 84 landmarks
Schmid et al. Schmid, Marx, and Samal (2008a)	78 variables: 6 for neoclassical canons, 55 for symmetry indicators, and 17 for golden ratios, obtained from 29 landmarks
Mao et al. Mao, Jin, and Du (2009a)	17 distances between landmarks
Sutic et al. Sutic, Breskovic, Huic, and Jukic (2010a)	25 geometric ratios derived from distances between 40 landmarks
Fan et al. J. Fan et al. (2012)	21 geometric ratios from 29 landmarks, reduced to 4 feature variables using PCA
Mu Mu (2013)	Normalized geometric coordinates of 36 facial landmarks
Dantcheva and Dugelay Dantcheva and Dugelay (2014a)	14 ratios of specified landmarks and one symmetry indicator
Liu et al. Shu Liu and Samal (June 2015)	11,651 frontal ratios from 82 frontal landmarks and 797 profile ratios from 40 profile landmarks
Xie et al. Xie, Liang, Jin, Xu, and Li (2015)	18 distances between landmarks
Liu et al. S. Liu, Fan, Guo, and Samal (2015)	124 principal components from 574 variables (318 ratios, 232 angles, and 24 inclinations from 82 frontal landmarks), and 51 principal components from 92 variables (24 ratios, 42 angles, and 26 inclinations from 40 profile landmarks)
Vahdati and Suen Vahdati and Suen (2021)	68 landmarks with 22 distances and 3 angles
Nezami and Suen Nezami and Suen (2023)	468 landmarks with 6 angles and 6 geometric features

The number of facial landmarks utilized for beauty prediction can differ significantly depending on the study, reflecting variations in methodology and the complexity of analysis. For example, (J. Fan et al., 2012) utilized 29 key facial landmarks, selecting them for their significant role in capturing general facial features. On the other hand, (S. Liu et al., 2015; Xie et al., 2015) employed

a more detailed approach, leveraging 82 frontal and 40 profile landmarks to provide a comprehensive representation of the facial structure, which potentially enhances the accuracy of attractiveness prediction by accounting for more nuanced facial features.

Beyond geometric features, researchers have explored additional hand-crafted features such as texture, skin smoothness, and color to refine the predictions of facial beauty models. These features help capture aspects of the face that are otherwise not easily represented by geometric landmarks alone. For instance, (Eisenthal et al., 2006b) employed features like hair color, facial symmetry, and indicators of skin smoothness to enhance their model's ability to assess beauty. (Kagian et al., 2008) used similar features, incorporating skin color, hair color, an asymmetry indicator, and a measure of skin smoothness. This combination allowed them to capture both structural and aesthetic aspects of attractiveness, thus offering a more holistic assessment. Dantcheva and Dugelay (Dantcheva & Dugelay, 2014a) went a step further by including eye color, recognizing that eye color, along with skin and hair, can significantly influence perceived attractiveness.

Moreover, Xie et al. (Xie et al., 2015) used Gabor features, which are particularly useful for capturing texture-related details, alongside geometric features to create a more robust representation of an individual's face. Gabor features help capture micro-level textures such as wrinkles and skin quality, contributing significantly to the model's ability to assess skin health and aesthetic appeal, both of which are important indicators of beauty.

Once these features are extracted, they are typically fed into traditional machine learning models, which are designed to predict facial attractiveness in a supervised manner. The features extracted are directly related to elements of facial attractiveness and are used to train models known as "auto-raters." The beauty labels assigned to the training data are derived from human judgments, thus aiming to create a model that mirrors subjective human perception. Depending on the problem definition, both classification and regression approaches have been used for these tasks, with varying degrees of success.

Regression methods, such as linear regression and support vector regression (SVR), are widely used to predict attractiveness scores. For instance, (Kagian et al., 2008) used a linear regression model to predict attractiveness and achieved a Pearson correlation of 0.82, indicating a strong relationship between their features and perceived beauty scores. Similarly, (Schmid et al., 2008a) used

linear regression and reported an R^2 value of 0.2433 with 78 predictor variables, which included features like symmetry and golden ratios. Although their results were somewhat lower, this outcome may have been influenced by the diversity of features or the complexity of human perception.

Interestingly, (J. Fan et al., 2012) removed confounding factors like hairstyle, skin texture, and facial expression from their dataset to achieve a clearer understanding of the intrinsic features that contribute to facial attractiveness. They reported an R^2 value of 0.6463, which is considerably higher than (Schmid et al., 2008a), despite the fact that both studies used largely similar features. This suggests that eliminating external factors may lead to more accurate models, as it allows the focus to remain on facial structure alone.

Other studies have experimented with more advanced regression techniques. (Mu, 2013), for instance, employed both Lasso and Ridge regression to handle the high-dimensional data, effectively reducing the risk of overfitting. (S. Liu et al., 2015) used SVR with an RBF kernel, a choice that enabled them to capture non-linear relationships between the features and attractiveness, achieving an R^2 value of 0.5756. In a comparative study, (Dantcheva & Dugelay, 2014a) utilized both linear regression and SVR, ultimately finding that linear regression performed better, with Pearson correlations of 0.65 and 0.557, respectively. This suggests that for some datasets, a simpler linear relationship may be sufficient to capture attractiveness.

Xie et al. (Xie et al., 2015) evaluated several regression techniques, including SVR, linear regression, pace regression, and Gaussian regression, focusing specifically on geometric features. They found that SVR had the highest performance, with a Pearson correlation of 0.608. Additionally, when they incorporated texture features along with geometric ones, the combined feature set further improved model performance, resulting in Pearson correlations of 0.6433 and 0.6482 for SVR and Gaussian regression, respectively. This result underscores the value of using a combination of geometric and texture-based features for more accurate beauty prediction.

In addition to the studies mentioned earlier, the work by (Vahdati & Suen, 2021) has made significant contributions to facial beauty prediction. Their approach utilizes 68 facial landmarks, from which 22 inter-landmark distances and 3 angles are derived. This methodology emphasizes the importance of both geometric distances and angular measurements in capturing the key structural features of the face that contribute to perceived attractiveness.

In addition to regression techniques, classification methods have also been employed to determine facial attractiveness. These approaches involve categorizing attractiveness into distinct classes rather than predicting a continuous score. Support vector machines (SVMs) and decision trees (C4.5) are two common classification techniques used in this context. (Mao et al., 2009a) applied both SVM and C4.5 classifiers and found that SVM provided better results, achieving accuracies of 77.9% for four-class classification and 95.3% for two-class classification. Similarly, (Sultan, Suen, Concordia University (Montréal, & Engineering, 2014) used SVM for classification and reported an impressive accuracy of 86%. These results highlight the effectiveness of SVM in capturing complex patterns within the data, making it a preferred choice for facial attractiveness classification.

Overall, the wide variety of features and machine learning techniques used in facial beauty prediction reflects the complexity of the task. Human attractiveness is influenced by numerous factors, including geometry, texture, color, and symmetry, and effective prediction requires careful feature selection and appropriate modeling techniques. Each study contributes to our understanding of facial beauty from different angles, and the continued exploration of both feature engineering and learning algorithms is crucial for advancing this field.

2.1.6 Deep Learning Models in facial Image Analysis

In recent years, cutting-edge deep learning techniques, particularly Convolutional Neural Networks (CNNs), have significantly advanced the accuracy and performance of facial beauty prediction models (L. Gao, Li, Huang, Huang, & Wang, 2018; Liang et al., 2018; Lin, Liang, & Jin, 2022; Shi, Gao, Meng, Xu, & Zhu, 2019; Shu Liu & Samal, June 2015; Vahdati & Suen, 2019; Xie et al., 2015; J. Xu, Jin, Liang, Feng, & Xie, 2015; L. Xu, Fan, & Xiang, 2019; L. Xu, Xiang, & Yuan, 2018a, 2018b). Unlike traditional approaches that rely on hand-crafted features, deep neural networks automatically learn high-level features directly from large datasets of facial images, resulting in more precise and scalable predictions.

Gan et al. (Gan, Li, Zhai, & Liu, 2014) introduced a deep self-taught learning approach to extract effective features for facial attractiveness prediction. Xie et al. (Xie et al., 2015) provided a benchmark dataset, SCUT-FBP, and developed a six-layer CNN designed to evaluate the attractiveness of female faces. Their model achieved a Pearson correlation of 0.8187, demonstrating the

superiority of CNN-extracted features over traditional geometric and Gabor features. This work highlighted the shift in facial beauty prediction toward deep learning-based methods.

Building on this, Xu et al. (Liang et al., 2018)developed a six-layer CNN using a cascaded fine-tuning approach, where multiple face input channels, including face images, detail layer images, and lighting layer images, were utilized. This multi-channel strategy improved the performance over Xie et al.'s model, with a Pearson correlation of 0.88, showing the advantages of integrating diverse facial data.

Fan et al. (L. Xu et al., 2019) took the field a step further by utilizing a deep convolutional residual network, ResNet, pre-trained on the ImageNet dataset (Deng et al., 2009). Their work introduced concepts such as label distribution learning (LDL) and feature fusion, which enriched their model. By incorporating these techniques, they achieved an impressive Pearson correlation of 0.93 on the SCUT-FBP dataset. The concept of label distribution learning was also employed by Gao et al. (B. Gao, Liu, Zhou, Wu, & Geng, 2020), who proposed a lightweight deep learning framework that integrated expectation regression modules. Despite having fewer parameters and faster inference speed, their model achieved nearly the same accuracy as Fan et al. (Y.-Y. Fan et al., 2018), proving that efficiency can be maintained without sacrificing performance.

Other works have explored different architectures. For example, Xu et al. (L. Xu et al., 2018b) used the VGG-Face model in combination with Bayesian ridge regression to assess facial beauty. By extracting deep features using a pre-trained VGG-16 network, they reported a Pearson correlation of 0.8570. Similarly, ResNet-18 was employed by another study (L. Xu et al., 2018a), which framed facial attractiveness prediction as both a regression and classification problem, yielding a Pearson correlation of 0.8723 through their CRNet model.

Further enhancing the field, Liang et al. (Liang et al., 2018) introduced the SCUT-FBP5500 dataset, a newer benchmark dataset for facial beauty prediction. They tested three CNN models pre-trained on ImageNet—AlexNet, ResNet-18, and ResNeXt-50—and found that ResNeXt-50 delivered the best performance, achieving a Pearson correlation of 0.8997. This result was later improved to 0.9142 by another study (Lin et al., 2022), where facial beauty prediction was formulated as a specific regression problem using ranking information to guide the model's learning process.

Shi et al. (Shi et al., 2019) introduced pixel-wise labeling masks along with a co-attention learning mechanism to improve the accuracy of attractiveness predictions, achieving a Pearson correlation of 0.926. Most recently, ComboLoss, a novel loss function combining regression, expectation, and classification losses, was proposed in (L. Xu & Xiang, 2020). Using the SEResNeXt50 network (Hu, Shen, & Sun, 2017), this approach delivered strong results with a Pearson correlation of 0.9199, illustrating the effectiveness of custom loss functions in beauty prediction tasks.

Lin et al. (Lin, Liang, Jin, & Chen, 2019) took a different route by introducing an Attribute-aware Convolutional Neural Network (AaNet), which incorporates beauty-related facial attributes as additional inputs to improve model performance. Their approach resulted in a Pearson correlation of 0.9055, highlighting the potential of combining attribute-based learning with deep CNN architectures.

2.1.7 Transformer-based models

In recent years, transformer-based architectures have brought about a paradigm shift in the field of computer vision, especially in image classification tasks. Among these, Vision Transformers (ViTs) have emerged as a powerful alternative to traditional convolutional neural networks (CNNs). By leveraging self-attention mechanisms, ViTs excel at capturing both global contextual relationships and subtle, fine-grained visual patterns within images. This capability makes them particularly well-suited for complex and subjective tasks such as facial beauty prediction, where nuanced visual cues play a critical role in determining perceived attractiveness.

2.1.8 Vision Transformer

Vision Transformers (ViTs) represent a significant departure from traditional convolutional neural networks (CNNs) by leveraging a self-attention mechanism originally introduced for natural language processing (NLP) tasks. First proposed by Dosovitskiy et al. (Dosovitskiy et al., 2020), ViTs divide an input image into a sequence of fixed-size patches, flatten and embed them, and then treat these embeddings as tokens—analogous to words in NLP. This patch-based approach allows ViTs to model long-range dependencies and contextual relationships across the entire image, without the spatial bias introduced by convolutional kernels.

This architectural innovation makes ViTs especially well-suited for tasks that require both fine-grained local detail and a holistic understanding of global structure—such as facial beauty prediction. Beauty assessment is inherently complex and subjective, often relying on subtle cues spread across the face, including symmetry, proportion, texture, and the interplay of features at various scales. ViTs' ability to attend to and integrate information from distant regions of the image enables more sophisticated modeling of such characteristics.

One of the key advantages of ViTs over CNNs lies in their capacity to capture global context efficiently. While CNNs are excellent at extracting localized features through hierarchical layers, they typically require deep stacks or additional mechanisms (e.g., dilated convolutions or global pooling) to capture global patterns. ViTs, on the other hand, naturally incorporate global attention from the earliest layers, making them more adept at evaluating features like facial symmetry and balance—elements that are crucial for assessing perceived attractiveness. Moreover, ViTs have shown strong generalization across diverse datasets, making them a compelling choice for robust and scalable beauty analysis systems.

2.1.9 Application in Facial Beauty Prediction

Vision Transformers (ViTs) have demonstrated exceptional performance across a wide range of image classification tasks, and their application in facial beauty prediction represents a highly promising direction for future research. These transformer-based models leverage self-attention mechanisms to dynamically focus on different regions of an image, enabling them to effectively capture both structural and aesthetic details of the face. This capacity for holistic and fine-grained analysis leads to more accurate, consistent, and interpretable predictions of perceived facial attractiveness.

In the context of beauty assessment, the strength of ViTs lies in their ability to model the complex interplay between key facial attributes—such as the geometric arrangement of facial landmarks, skin texture, and overall symmetry and proportion. These characteristics are closely tied to human perceptions of beauty, yet traditional CNN-based methods often struggle to capture such global and nuanced relationships due to their inherently localized receptive fields. ViTs, by contrast, can aggregate contextual information across the entire facial image, resulting in a more comprehensive

and perceptive feature representation.

Moreover, the inherent adaptability of transformer-based models makes them more robust to real-world variability, including changes in lighting conditions, facial expressions, and head orientations. This robustness enhances their generalization capabilities across diverse datasets and demographic groups, addressing key challenges in the development of fair and reliable facial beauty prediction systems.

As deep learning architectures continue to advance, the integration of ViTs into facial beauty prediction pipelines is expected to further improve performance, particularly in terms of accuracy, generalization, and interpretability. Their ability to learn from large-scale, diverse datasets while maintaining transparency in decision-making offers new opportunities for promoting fairness and inclusivity—supporting broader recognition of cultural and individual differences in aesthetic standards.

Chapter 3

Dataset

Building upon the studies referenced in (Laurentini & Bottino, 2014; S. Liu et al., 2016), this chapter provides a comprehensive review of face datasets utilized in existing research, analyzing them based on attributes such as size, gender, ethnicity, age, pose, expression, data sources, and the characteristics of human raters. Furthermore, a publicly available benchmark dataset, (Liang et al., 2018), which serve as the foundation for this study, is introduced.

3.1 Face Beauty Datasets

The selection of an appropriate face dataset plays a crucial role in the analysis of facial beauty. The dataset used can significantly influence the accuracy and reliability of the predicted beauty scores, making the choice of dataset a critical factor in research outcomes. An ideal face dataset should encompass a diverse range of attributes, including varying levels of attractiveness, a wide age spectrum, representation of multiple genders, and individuals from different ethnic backgrounds. This diversity ensures that the analysis captures a more comprehensive and unbiased understanding of facial beauty. As emphasized in (S. Liu et al., 2016), employing datasets with such varied characteristics is essential for achieving meaningful and generalizable results in facial beauty prediction studies.

3.1.1 Collecting Facial Images

Face images can be obtained from a wide array of sources, each contributing unique characteristics and diversity to the dataset. These sources include publicly available face datasets, which often serve as benchmarks in facial analysis research, as well as images retrieved from the Internet, which provide access to a vast pool of real-world data. Additionally, computer-generated visuals offer a controlled and scalable means of generating synthetic data, allowing researchers to study facial features under varying conditions. Photographs and face models captured using digital cameras or advanced 3D scanners further enrich datasets by providing high-quality and detailed representations of facial structures (S. Liu et al., 2016). The use of such varied sources ensures that datasets are robust and capable of addressing the complex requirements of facial beauty analysis.

Tables 3.1 and 3.2 present a detailed breakdown of the attributes associated with face datasets utilized in existing studies. These attributes include the number of face images available in each dataset, the age range of the individuals represented, the gender distribution, and the ethnic diversity of the subjects. Additionally, the datasets are categorized based on pose variations, facial expressions, and the specific sources from which the images were obtained. This information provides researchers with critical insights into the composition and suitability of datasets for various applications. For readers interested in an in-depth exploration of face attractiveness research, references (Laurentini & Bottino, 2014) and (S. Liu et al., 2016) serve as valuable resources, offering comprehensive reviews and discussions on the methodologies and findings in this evolving field. These reviews highlight the significance of dataset selection in advancing facial beauty prediction and other related research areas.

Table 3.1: Review of the face datasets used in existing works in terms of size, gender (F and M indicate female and male), face ethnicity and age.

Reference	**Size**	**Gender**	**Ethnicity**	**Age**
(Gunes & Piccardi, 2006)	215	F	diverse	different
(Eisenthal, Dror, & Ruppin, 2006c)	Set1:92, Set2:92	F	American, Israeli	young, about 18
(Kagian et al., 2008)	91	F	American	young
(Schmid et al., 2008a)	452	F/M	Caucasian	N/A
(Mao, Jin, & Du, 2009b)	510	F	Chinese	N/A
(Sutic, Breskovic, Huic, & Jukic, 2010b)	(136), (200)	F	diverse	young
(J. Fan et al., 2012))	545	F	N/A	N/A
(Rizvi & Karawia, 2013)	30	F	diverse	young
(Mu, 2013)	250	F/M	Asian	20-40
(Dantcheva & Dugelay, 2014b)	325	F	diverse	young
(S. Liu et al., 2017a)	180	F	Chinese	16-49
(Xie et al., 2015)	500	F	Asian	young
(S. Liu et al., 2017b)	360	F/M	Chinese	16-49
(Liang et al., 2018)	5500	F/M	Asian, Caucasian	15-60

Table 3.2: Review of the face datasets used in existing works in terms of pose, expression and sources.

Reference	**Pose**	**Expression**	**Source**
(Gunes & Piccardi, 2006)	frontal	N/A	captured
(Eisenthal et al., 2006c)	Set1: frontal, Set2: frontal	neutral, almost neutral	captured
(Kagian et al., 2008)	frontal	neutral	captured
(Schmid et al., 2008a)	frontal	neutral	FERET database, Internet
(Mao et al., 2009b)	frontal	neutral	Internet, CAS-PEAL face dataset
(Sutic et al., 2010b)	frontal	different	Internet
(J. Fan et al., 2012)	frontal	neutral	synthesized
(Rizvi & Karawia, 2013)	frontal	almost neutral	Internet
(Mu, 2013)	frontal	almost neutral	Internet
(Dantcheva & Dugelay, 2014b)	almost frontal	different	Internet
(S. Liu et al., 2017a)	frontal, lateral	neutral	BJUT-3D face database
(Xie et al., 2015)	frontal	neutral	captured, Internet
(S. Liu et al., 2017b)	frontal, lateral	neutral	BJUT-3D face database
(Liang et al., 2018)	frontal	neutral	Internet, 10k US Adult Face database

3.2 Characteristics of Face Datasets in Facial Beauty Analysis

3.2.1 Face Dataset Size:

The size of face beauty datasets varies significantly, ranging from as few as 30 images in some studies (Rizvi & Karawia, 2013) to as many as 5,500 images in larger datasets (Liang et al., 2018). This variation in dataset size plays a crucial role in determining the robustness and generalizability of the models used for facial beauty prediction.

3.2.2 Face Gender:

The majority of facial beauty datasets predominantly feature images of female faces. This gender imbalance highlights a limitation in dataset diversity, which could impact the fairness and inclusivity of models trained on such data.

3.2.3 Face Age and Ethniciy:

Diverse ethnic backgrounds have been represented in several research studies (Dantcheva & Dugelay, 2014b; Gunes & Piccardi, 2006; Rizvi & Karawia, 2013; Sutic et al., 2010a). Some datasets focus exclusively on specific ethnic groups, such as Asian faces (Xie et al., 2015) or Caucasian faces (Schmid et al., 2008a), while others incorporate a mix of ethnicities. Regarding age, a wide age range is considered in certain datasets, such as (Liang et al., 2018), providing a more comprehensive view of facial beauty across different life stages. However, other datasets focus solely on young faces, limiting the scope of analysis (Kagian et al., 2008; Sutic et al., 2010a; Xie et al., 2015).

3.2.4 Face Pose:

Many research studies concentrate on frontal-view faces, as this pose is often deemed most suitable for analyzing facial features. Notably, Liu et al. (S. Liu et al., 2015, 2017b) expanded this scope by including both frontal and profile-view faces in their studies, introducing greater variability in pose and enriching the dataset for more diverse analysis.

3.2.5 Facial Expression:

A significant portion of facial beauty research is based on datasets containing faces with neutral expressions. This preference simplifies the analysis and ensures consistency in evaluating facial features. However, it also leaves room for exploring the impact of various facial expressions on perceived beauty.

These characteristics collectively illustrate the diversity and limitations of existing face beauty datasets, providing valuable insights into how dataset attributes influence research outcomes and highlighting areas for future improvement.

3.3 Pre-processing

Face images are typically collected from various sources and undergo pre-processing to prepare them for analysis. Techniques such as face cropping and landmark localization are critical steps to ensure that the images are standardized and suitable for model training and evaluation [59]. These processes help in isolating the facial regions of interest and aligning features for consistent analysis. It is worth noting that some public face datasets, like the SCUT-FBP5500 dataset [53], come pre-processed, reducing the workload for researchers and providing ready-to-use data for studies.

3.3.1 Attractiveness Score Collection

One of the significant challenges in facial beauty analysis is the absence of universally agreed-upon true beauty scores (S. Liu et al., 2016). To address this, researchers rely on human raters to provide attractiveness scores for face images. This process involves asking human raters to evaluate facial attractiveness on a defined scale, typically using integer values to quantify their assessments. For instance, in a 7-point scale system, raters assign scores where 1 represents the least attractive and 7 represents the most attractive. The ground truth for each image is then calculated as the average or median of the scores assigned by the raters.

A detailed summary of the characteristics of human raters is presented in Table 5. Notably, the number of participants varies significantly across studies. For example, 28 raters were employed

in the works of (Eisenthal et al., 2006a; Kagian et al., 2008), while over 70 participants were involved in the study conducted by (Dantcheva & Dugelay, 2014a). Most studies ensure diversity by employing both male (M) and female (F) raters. Furthermore, some research incorporates raters of different ages and ethnic backgrounds to provide a more balanced perspective on facial beauty (Dantcheva & Dugelay, 2014a; Gunes & Piccardi, 2006; Shu Liu & Samal, June 2015; Sutic et al., 2010a). However, other studies use a more homogeneous group of judges, as seen in [24].

The scoring scales used by human raters also differ across studies, ranging from a 4-point scale (Mao et al., 2009b) to a 10-point scale (Dantcheva & Dugelay, 2014b; Gunes & Piccardi, 2006; S. Liu et al., 2015, 2017b; Schmid et al., 2008a; Sutic et al., 2010a, 2010b). This variation in scoring systems reflects the diverse methodologies employed in facial beauty analysis and highlights the need for standardization to enable consistent comparisons across studies.

Table 3.3: A summary of human raters' characteristics.

Reference	**Number**	**Gender**	**Ethnicity**	**Age**	**Rating**
(Gunes & Piccardi, 2006)	48	F/M	diverse	Above 18	10-point scale
(Eisenthal et al., 2006c)	28 for set1, 18 for set2	F/M	N/A	20-29	7-point scale
(Kagian et al., 2008)	28	F/M	N/A	N/A	7-point scale
(Schmid et al., 2008a)	36	F/M	N/A	19-61	10-point scale
(Mao et al., 2009b)	N/A	N/A	N/A	20-29	4-point scale
(Sutic et al., 2010b)	¿50	F/M	diverse	different	10-point scale
(J. Fan et al., 2012)	30	F/M	Chinese	21-27	9-point scale
(Rizvi & Karawia, 2013)	29	N/A	N/A	N/A	5-point scale
(Mu, 2013)	32	F/M	N/A	20-30	7-point scale
(Dantcheva & Dugelay, 2014b)	less than 70	F/M	diverse	different	10-point scale
(S. Liu et al., 2017a)	48	F/M	diverse	19-76	10-point scale
(Xie et al., 2015)	about 70	F/M	N/A	N/A	5-point scale
(S. Liu et al., 2017b)	48	F/M	diverse	N/A	10-point scale
(Liang et al., 2018)	60	F/M	N/A	18-27	5-point scale

3.4 SCUT-FBP5500 Dataset

This dataset comprises a total of 5,500 high-resolution frontal face images, representing individuals aged between 15 and 60 years, all captured with neutral facial expressions. As shown in Figure 12, the dataset is structured into four distinct subsets based on race and gender, including 2,000 Asian females, 2,000 Asian males, 750 Caucasian females, and 750 Caucasian males. The majority of these facial images were sourced from the Internet, ensuring a diverse range of features and appearances. Additionally, a portion of the Asian face images was obtained from the DataTang2 collection, while some Caucasian face images were sourced from the 10k US Adult Face database (Bainbridge, Isola, & Oliva, 2013). This curated combination of images makes the dataset a valuable resource for studies requiring balanced representation across genders and ethnic groups.

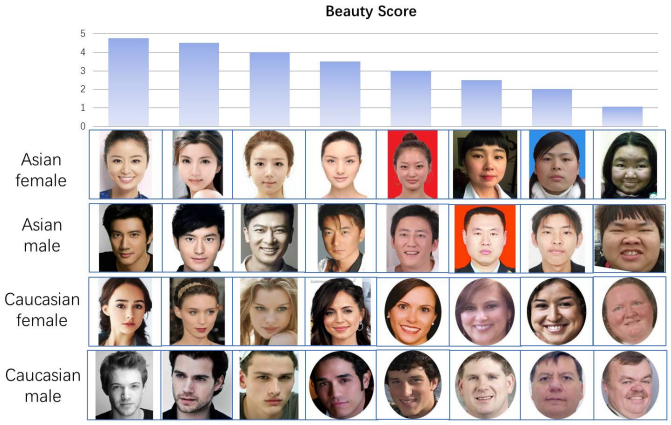

Figure 3.1: SCUT-FBP5500 dataset overview (sourced from (Liang et al., 2018)).

The face images in this dataset have been meticulously labeled by 60 human raters, aged between 18 and 27 years, with an average age of 21.6. Each facial image was rated using integer scores on a scale from 1 to 5, where a score of ”1” represents the least attractive and ”5” indicates extremely attractive. To minimize variations in the labeling process, approximately 10% of the face images were randomly repeated during the rating sessions. If the correlation coefficient between the two attractiveness scores of the same image was found to be less than 0.7, the rater was asked to reassess the image to ensure consistent and reliable ratings.

The final attractiveness score for each face image was determined by averaging the scores provided by all 60 human raters, thus creating a robust ground truth (beauty label) for subsequent analyses.

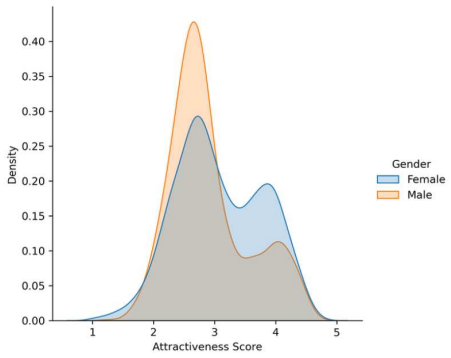

Figure 3.2: SCUT-FBP5500 dataset overview (sourced from (Vahdati & Suen, 2021)).

Chapter 4

Methodology

One of the core challenges in facial attractiveness prediction is the extraction of meaningful and accurate facial representations. Traditionally, many studies have relied on hand-crafted features—particularly geometric ones—to quantify facial aesthetics. While such features may encode some structural information, they are inherently limited in their ability to capture the subtle and complex traits that contribute to human perceptions of beauty. Consequently, shallow models that depend on these heuristics are prone to under performance, restricting the potential of attractiveness prediction.

With the advent of deep learning, Convolutional Neural Networks (CNNs) have revolutionized facial analysis tasks by enabling automatic extraction of high-level, hierarchical features from raw images (Y.-Y. Fan et al., 2018; L. Gao et al., 2018; Liang et al., 2018; Lin et al., 2022; Shi et al., 2019; Xie et al., 2015; J. Xu et al., 2015; L. Xu et al., 2019, 2018a). These models learn discriminative representations that are critical for face-related tasks, outperforming traditional approaches by a substantial margin. However, despite these advances, existing deep models for facial attractiveness prediction often face challenges such as limited robustness when trained on small datasets, susceptibility to bias in demographic attributes, and a lack of interpretability—factors that can limit their reliability in practical applications.

In this chapter, we propose a novel framework for predicting facial beauty scores using a transformer-based architecture combined with transfer learning. Specifically, we adopt a Vision Transformer (ViT) model that was initially pretrained on large-scale datasets for face recognition

tasks such as VGGFace2 (Cao, Shen, Xie, Parkhi, & Zisserman, 2017), Glint360K (An et al., 2022), MS1MV2 (Guo, Zhang, Hu, He, & Gao, 2016). We then fine-tune this model on our target dataset, SCUT-FBP5500—a benchmark facial beauty dataset containing 5,500 images annotated with attractiveness scores from multiple raters and labeled for gender and ethnicity—to adapt it for facial attractiveness prediction. The fine-tuned model extracts deep facial features, which are subsequently passed to a regression head to predict a continuous beauty score for each individual.

In addition, we extend this methodology to build a complete multi-task framework capable of predicting not only beauty scores but also gender and ethnicity. This design enables our model to simultaneously address multiple facial analysis tasks, enhancing its robustness and applicability.

Our approach for facial beauty prediction, which is our main challenge, comprises two main stages:

1- Feature Extraction: We begin with a Vision Transformer (ViT) model pretrained on a large-scale face recognition datasets. This model is then fine-tuned on the SCUT-FBP5500 dataset to adapt it specifically for facial beauty analysis. Through this process, the model learns to extract high-level, task-specific representations from face images.

2- Prediction: The fine-tuned ViT model outputs feature embeddings, which are passed through a regression head to predict a continuous beauty score for each individual.

4.1 Data Pre-processing and Augmentation approaches

To improve robustness and enhance the generalization capability of the model, data pre-processing and augmentation were applied primarily to the training set. All images were first resized to 224×224 pixels to match the input size of the Vision Transformer (ViT) model. Pixel values were normalized using the mean and standard deviation of the ImageNet dataset, consistent with the pretraining configuration.

Standard online augmentation techniques included random rotations (up to $\pm 15^{\circ}$) and random horizontal flipping with a probability of 0.4. These augmentations introduce controlled variability in facial orientation and symmetry, reducing overfitting to specific poses present in the training data.

In addition to standard augmentations, a custom alignment step was applied to preserve bilateral facial symmetry. Facial landmarks were used to extract the coordinates of the left and right eye, which were then used to estimate the head tilt angle. An affine transformation was applied to rotate each image so that the eyes were horizontally aligned. This alignment ensures that key symmetric facial features—particularly the eyes—are consistently positioned across samples. By minimizing pose variation and eliminating artifacts caused by head tilts, the model focuses on genuine facial attributes rather than orientation-specific patterns, thereby improving prediction stability and accuracy across all tasks.

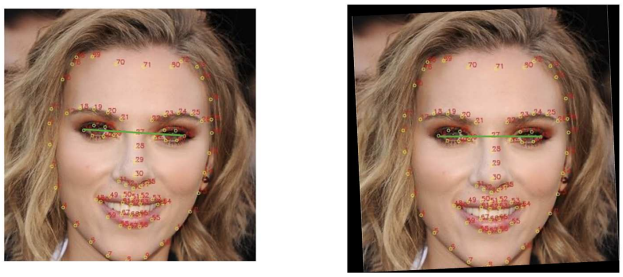

Figure 4.1: A sample of the custom augmentation procedure for aligning facial symmetry.

4.2 Model Selection and Feature Learning Pipeline

4.2.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are widely used in computer vision due to their ability to learn hierarchical image features from raw pixel data. A typical CNN consists of convolutional layers for spatial feature extraction, pooling layers for downsampling, and fully connected layers for the final prediction stage. For regression problems such as facial beauty prediction, the output layer produces a continuous value rather than a discrete class label.

In CNNs, early layers capture low-level visual features such as edges, textures, and color gradients, while deeper layers encode higher-level semantic information, including complex facial structures. This hierarchical representation learning has made CNNs highly effective for face-related tasks, including expression recognition, identity verification, and attractiveness prediction.

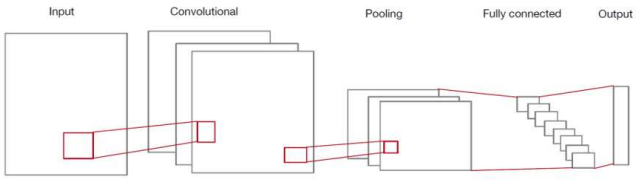

Figure 4.2: General architecture of a CNN illustrating convolutional, pooling, and fully connected layers, as used in facial analysis tasks

When training data is limited, transfer learning with pre-trained CNNs is a common strategy. By initializing the model with weights learned from large-scale datasets, the network benefits from a strong starting point, enabling faster convergence and improved generalization to the target task.

4.2.2 Residual Neural Networks

Residual Networks (ResNets), proposed by He et al. (He et al., 2015), introduced skip connections to address the vanishing gradient problem in deep neural networks. In a residual block, the output is computed as:

$$\mathcal{H}(\mathbf{x}) = \mathcal{F}(\mathbf{x}) + \mathbf{x}$$

where $\mathcal{F}(\mathbf{x})$ is the residual function learned by the network.

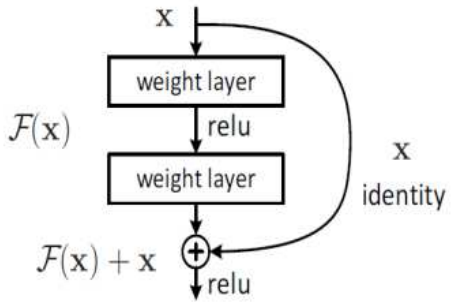

Figure 4.3: A residual block illustrating a shortcut connection between layers, enabling residual learning and mitigating the vanishing gradient problem (He et al., 2015).

These shortcut connections allow gradients to propagate more effectively during backpropagation, enabling the training of much deeper architectures without performance degradation. ResNet

architectures have achieved state-of-the-art performance across numerous computer vision benchmarks and are frequently used as feature extractors in face-related tasks, including identity recognition, expression classification, and attractiveness prediction. In our study, ResNets serve as a conceptual foundation for understanding deep feature extraction prior to exploring transformer-based models.

4.2.3 Transfer Learning and Deep Feature Extraction

Transfer learning is particularly effective when the available training data is limited, as it enables models to leverage knowledge acquired from large-scale datasets in related domains. In this work, we employ a Vision Transformer (ViT) backbone pretrained on large-scale face recognition datasets to extract meaningful features from our comparatively small dataset.

Two common strategies for transfer learning are:

Fixed feature extractor: Only the final prediction head is trained, while the backbone parameters remain frozen.

Fine-tuning: The pretrained network is partially or fully retrained on the new data.

Given that our dataset is relatively small but closely related in domain to the pretraining data (human faces), we adopt a two-stage transfer learning strategy. First, a ViT model is initialized from ImageNet pretraining and further trained on large-scale face recognition datasets such as VGGFace2, Glint360K, and MS1MV2. This intermediate step provides domain-specific facial feature representations, reducing the domain gap between the source and target tasks.

In the second stage, all layers of the model are fine-tuned on the SCUT-FBP5500 dataset for facial attractiveness prediction. This approach enables the network to adapt its learned facial representations to the subtle, subjective characteristics of beauty assessment while retaining the generalization capabilities gained from large-scale pretraining. The following section describes the Vision Transformer architecture in detail.

4.2.4 Vision Transformer (ViT)

Before describing the Vision Transformer (ViT) used in this study, we briefly review the general Transformer architecture, including both encoder and decoder components, as originally proposed

by (Vaswani et al., 2017) to understanding the full architecture because it provides useful context for its design principles.

"Attention Is All You Need"

The transformer architecture, originally proposed for machine translation tasks (Vaswani et al., 2017), consists of encoder and decoder stacks, leveraging self-attention to model global dependencies in data sequences. In the ViT architecture, only the encoder is used to process visual inputs.

The encoding component comprises a stack of encoders, while the decoding component is a stack of decoders, both with the same number (six as per the original paper). All encoders and decoders share an identical structure, though their weights are distinct.

Each one consists of two and three sub-layers, respectively. Now, we discuss how self-attention and encoder-decoder attention work.

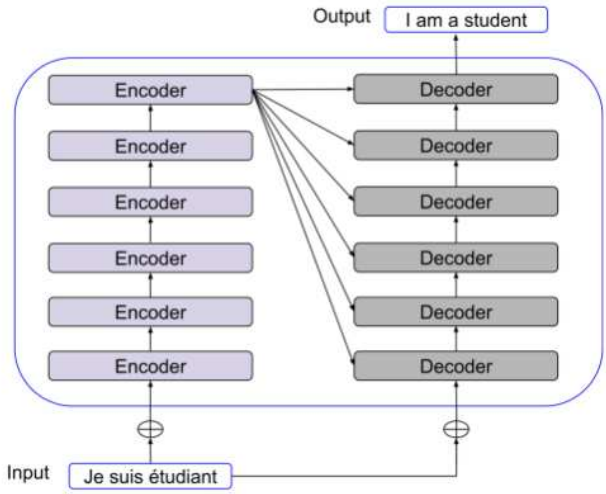

Figure 4.4: An example of an encoder-decoder for a translation task

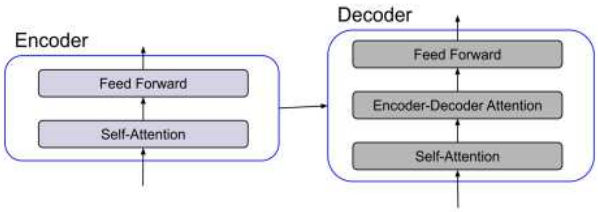

Figure 4.5: Encoder-decoder in ViT (NLP)

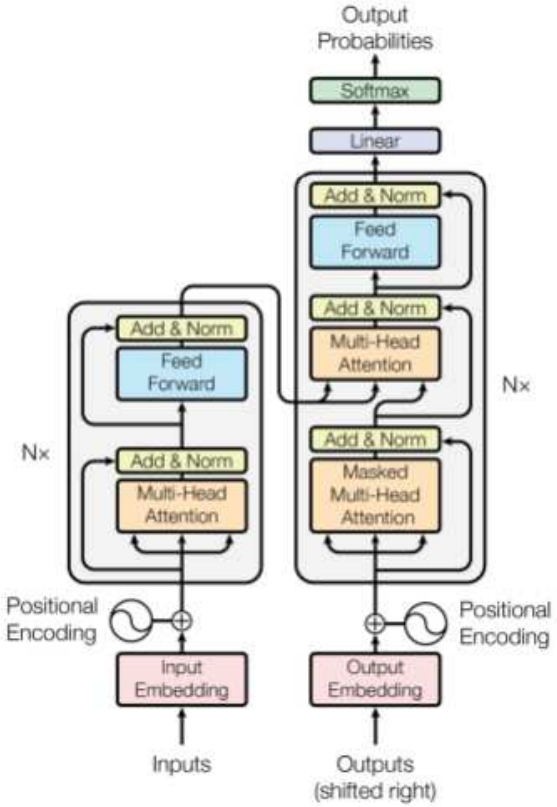

Figure 4.6: The transformer architecture (Image is taken from (Vaswani et al., 2017))

In addition to the encoder and decoder components, the transformer architecture requires the use of word embeddings. Embeddings are applied only at the bottom-most encoder and decoder layers. The standard convention is that each encoder receives a list of vectors, with each vector having the same dimensionality as the model's projection dimension. In the lowest encoder (and

decoder), these vectors are derived from word embeddings, whereas in the subsequent layers, they represent the output of the layer directly below.

A key mechanism in the transformer is self-attention, which allows the model to attend to different positions within the input sequence, capturing contextual information relevant to each element. To achieve this, the model computes three matrices: Query (Q), Key (K), and Value (V), which are derived by multiplying the input embeddings (represented as a matrix X) with learned weight matrices W_Q, W_K, and W_V, respectively:

$$\begin{aligned} Q &= XW^Q \\ K &= XW^K \\ V &= XW^V \end{aligned} \tag{1}$$

Given the Query (Q), Key (K), and Value (V) matrices computed from the input matrix X (which may represent, for example, three-word embeddings), the self-attention output is calculated using the following formula:

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

In the original transformer architecture, the dimension of the key vectors d_k is set to 64. The concept of **multi-head attention** involves repeating the self-attention mechanism multiple times in parallel. Specifically, the original paper introduces eight parallel self-attention heads, each with its own set of learned parameters: $W_Q^{(i)}$, $W_K^{(i)}$, and $W_V^{(i)}$, where i denotes the head index.

As a result, the input sequence produces eight distinct output representations—one from each attention head—commonly referred to as Z_i. These output vectors from all heads are concatenated along the feature dimension to form a single matrix. This concatenated matrix is then projected using a final learnable linear transformation matrix W_O, yielding the final output of the multi-head attention mechanism.

$$Z = [Z_0 \quad Z_1 \quad Z_2 \quad Z_3 \quad Z_4 \quad Z_5 \quad Z_6 \quad Z_7]W^O \tag{3}$$

The output of the multi-head attention mechanism is denoted as matrix Z. This matrix undergoes a **residual connection** (addition of the original input) followed by **layer normalization**. The resulting normalized output is then passed through a **position-wise feedforward network**, which typically consists of two linear layers with a non-linear activation function in between. This is followed by another residual connection and layer normalization step. These operations—multi-head attention, addition and normalization, and the feedforward network—constitute the key components of each encoder block. Once these steps are completed, the final output of the encoder stack is ready to be passed into the decoder for further processing.

Now, we turn our attention to the decoder component of the transformer architecture. Similar to the encoder, the decoder is initially provided with a sequence of word embeddings. However, during training, it is essential to prevent the decoder from accessing future tokens in the target sequence—a process known as **causal masking**. To implement this, a **masking mechanism** is applied to the self-attention computation. Specifically, a **mask matrix**—an upper triangular matrix with values of $-\infty$ above the main diagonal—is added to the attention score matrix QK^T. This ensures that the softmax operation assigns zero probability to any "future" tokens, effectively blocking the decoder from attending to information beyond the current position.

The output of this **masked multi-head attention** layer then undergoes a **residual connection and layer normalization**. Following this, the result is passed to another multi-head attention mechanism, which incorporates the **output of the encoder stack**. In this second attention layer, the Query (Q) is computed from the decoder, while the Key (K) and Value (V) matrices are obtained from the final encoder outputs via linear projections.

This cross-attention output also undergoes addition and normalization, followed by a **position-wise feedforward network**, another residual connection, and final normalization. The resulting decoder output is passed through a linear layer and a softmax function to compute the **probability distribution over the target vocabulary**. The word with the highest probability is selected as the predicted output token.

While the encoder processes the input sequence only once, the decoder operates **autoregressively**—predicting one token at a time—until an end-of-sequence (<eos>) token is generated, indicating the completion of the predicted sequence (Mahmud, 2020).

4.2.5 An image is worth 16x16 words

The Vision Transformer (ViT) utilizes the encoding component of the transformer model. In this approach, an image is divided into non-overlapping, fixed-size patches, each of size $P \times P$. These patches are then flattened and linearly transformed using a trainable matrix E.

$$patch\_embedding = flattened\_patch \cdot E \tag{4}$$

In Vision Transformer architectures, common patch sizes include $P = 16$ and $P = 32$. The choice of patch size presents a trade-off. Larger patch sizes reduce the number of patches extracted from an image, which decreases the model's ability to capture fine-grained spatial relationships between local regions. However, they also reduce computational cost by lowering the sequence length passed to the transformer. Conversely, smaller patch sizes preserve more spatial detail but increase memory and compute requirements due to the larger number of patches. Ultimately, the optimal patch size depends on empirical evaluation and the specific characteristics of the target task. After the patches are generated, **positional encodings** are added to each patch embedding to retain information about spatial order. The sum of the patch embedding and its positional encoding is then fed into the transformer encoder for further processing.

$$embedding\_vector = patch\_embedding + positional\_encoding \tag{5}$$

Additionally, unlike natural language processing tasks where a special start token is typically used, the Vision Transformer (ViT) architecture introduces a **learnable class token**, commonly referred to as **'patch0'**. This token is prepended to the sequence of patch embeddings and is designed to aggregate information from all other patches during the self-attention process. The final output from the transformer encoder includes contextualized embeddings for each input patch, with **'patch0'** serving as a global representation of the entire image. Due to its role in summarizing the full visual context, **'patch0'** holds particular importance in downstream tasks. In our framework, we focus exclusively on the **'patch0'** output and pass it through a **Multi-Layer Perceptron (MLP) head**, which is responsible for making the final prediction. While the original ViT model may use

this architecture for classification tasks such as distinguishing between several cases, in our case, this head is adapted to predict a continuous **beauty score**.

4.2.6 Our ViT Model

In our proposed framework, we employ a customized Vision Transformer (ViT) architecture specifically adapted for facial beauty prediction. The model initialization follows a **two-stage transfer learning strategy**:

(1) **Stage 1 – Domain-specific pretraining:** We start with a ViT model pretrained on ImageNet and further trained on the large-scale VGGFace2 dataset for face recognition. This step enables the model to learn rich and diverse facial representations that capture structural and identity-related features.

(2) **Stage 2 – Task-specific fine-tuning:** All layers of the pretrained model are fine-tuned on the SCUT-FBP5500 dataset for attractiveness prediction, allowing the model to adapt its learned facial features to the subtle and subjective characteristics of beauty assessment.

Input images are resized to 224×224 pixels and divided into non-overlapping patches of 16×16 or 32×32 pixels, depending on the experiment. Each patch is flattened and projected into a latent embedding space via a learnable linear transformation, with positional encodings added to preserve spatial relationships. A learnable **class token** (CLS, referred to in the implementation as `patch_0`) is prepended to the patch sequence, and its final encoded representation—after passing through the transformer encoder layers—serves as a global embedding summarizing the entire image.

For baseline attractiveness prediction, this CLS token embedding is passed through a task-specific Multi-Layer Perceptron (MLP) head to output a single continuous beauty score. The next subsection describes how this single-task baseline is extended into a multi-task learning framework to jointly predict additional facial attributes.

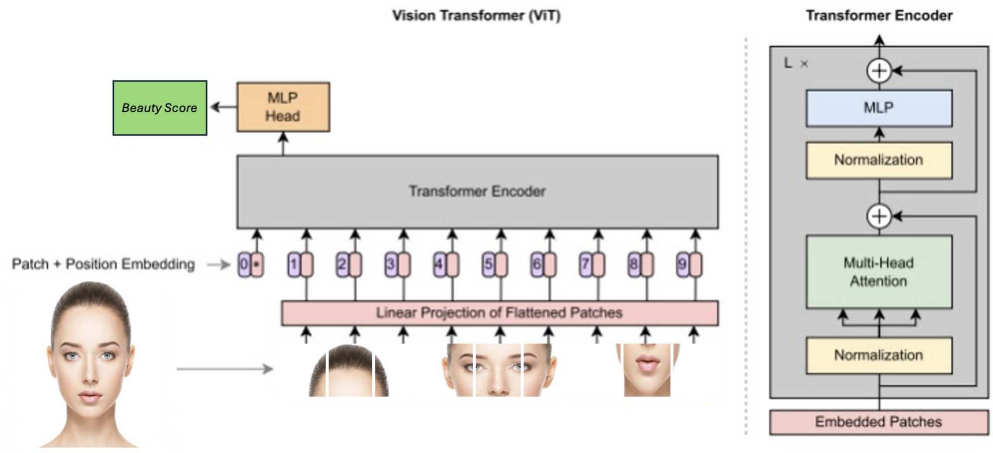

Figure 4.7: Our Vision Transformer framework for attractiveness prediction

4.2.7 Multi-task Learning Scheme

To enrich the attractiveness prediction task with additional contextual cues, we employ a multi-task learning (MTL) framework. In this approach, facial attractiveness is formulated as a regression problem, while both gender and ethnicity are treated as binary classification problems. The key motivation for adopting this setup is to leverage shared facial representations for multiple related tasks, allowing the model to learn more robust and generalizable features that capture both aesthetic and demographic attributes.

The proposed architecture is built upon a pretrained Vision Transformer (ViT) backbone, which serves as a shared feature extraction module. Input images are processed through the ViT, and the global embedding derived from the [CLS] token is used as a common representation for all tasks. This embedding is then passed to three independent task-specific prediction heads. The beauty regression head outputs a single continuous score representing perceived attractiveness. The gender classification head produces a probability value through a sigmoid activation, indicating the likelihood of the face being male or female. Similarly, the ethnicity classification head outputs a sigmoid-activated probability, with class labels encoded as 0 (A) and 1 (C).

Each task-specific head is composed of fully connected layers and may include dropout for

regularization to reduce overfitting. The design ensures that while the feature extractor is shared across all tasks, each head can learn its own decision boundaries tailored to its objective.

The training objective is defined as a weighted sum of the individual task losses:

$$\mathcal{L}\text{total} = \lambda_B \mathcal{L}\text{MSE}^B + \lambda_G \mathcal{L}\text{BCE}^G + \lambda_E \mathcal{L}\text{BCE}^E, \tag{6}$$

where $\mathcal{L}\text{MSE}^B$ is the Mean Squared Error loss for attractiveness regression, and $\mathcal{L}\text{BCE}^G$ and $\mathcal{L}_{\text{BCE}}^E$ are the Binary Cross-Entropy losses for gender and ethnicity classification, respectively. The coefficients λ_B, λ_G, and λ_E control the relative importance of each task in the overall optimization process. By jointly optimizing these objectives, the model is encouraged to learn a shared representation that improves performance across all tasks, while still allowing each head to specialize in its specific prediction target.

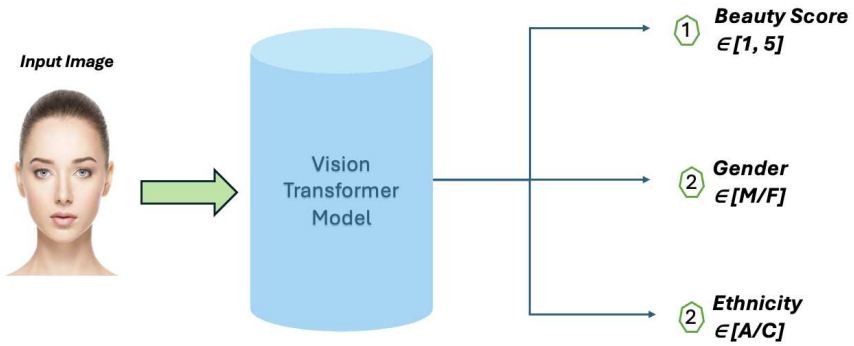

Figure 4.8: Proposed Vision Transformer multi-task framework for joint prediction of facial attractiveness (regression), gender (binary classification), and ethnicity (binary classification).

Chapter 5

Experiments and Results

This chapter describes the experimental setup, evaluation protocols, and results obtained from testing multiple variants of the ViT architecture. Our experiments are designed to evaluate the impact of different configurations on model performance and to analyze how architectural capacity interacts with limited dataset size and the risk of overfitting.

Table 5.1 summarizes the Vision Transformer (ViT) variants under investigation, ordered by their parameter counts. These results highlight a substantial variation in model size across configurations. Unlike many convolutional architectures, transformers adopt weaker inductive biases, enabling them to capture long-range dependencies without strong prior assumptions. While this flexibility leads to higher representational capacity, it also comes with an increased susceptibility to overfitting when training data is limited.

Table 5.1: Comparison of model parameter counts for ViT variants.

Model	**Parameters**
ViT-Large	307M
ViT-Base	86M
ViT-Small	22.2M
ViT-Tiny	8M

Models with a greater number of learnable parameters generally exhibit an increased capacity to capture and represent complex patterns in the input data. However, this comes at the cost of a heightened risk of overfitting, particularly when the available dataset is limited in size. To mitigate this challenge, we employed data augmentation techniques to effectively expand the training set

and promote generalization. Additionally, a higher dropout rate was used to reduce co-adaptation of neurons, and the training process was closely monitored using loss curves and performance metrics recorded at each epoch.

For our facial attractiveness prediction framework, we adopted Vision Transformer models pretrained on face recognition tasks. This strategy provides a strong initialization, as the models have already learned rich and discriminative facial representations during large-scale pretraining. Specifically, our best-performing ViT backbone was pretrained (Rodrigo, Cuevas, & García, 2024) on the VGGFace2 dataset — a benchmark comprising over 3.3 million images from more than 9,000 unique identities, covering a wide range of variation in pose, age, ethnicity, and illumination.

The base ViT configurations operate using fixed input resolutions of either 224×224 or 384×384 pixels, following their original training protocol. To ensure compatibility with the pretrained weights and preserve spatial alignment of features, all images in our dataset were resized to comply with the corresponding input resolution. This preprocessing step standardizes the data pipeline and enables fair performance comparisons across different ViT configurations.

5.0.1 Experimental Setup

All experiments were conducted on a high-performance workstation equipped with an Intel Core i9-13900K CPU, two NVIDIA RTX 4090 GPUs, and 128 GB of RAM, running Ubuntu 22.04 LTS. The dual-GPU configuration enabled efficient parallel computation and reduced training time for large-scale transformer models.

All models were implemented in Python using the PyTorch framework. To maximize hardware utilization, we employed data-parallel training, splitting each mini-batch across both GPUs. This configuration allowed for larger effective batch sizes without exceeding GPU memory limits. The facial attractiveness models were trained for up to 150 epochs (typically converging within 100–150), with early stopping based on validation performance.

Several optimization strategies were explored during preliminary experiments. The Adam optimizer was ultimately selected due to its adaptive learning-rate properties, which are particularly effective for transformer-based architectures. A short linear warm-up phase was followed by a `ReduceLROnPlateau` scheduler, which reduced the learning rate adaptively when validation

performance plateaued. For the attractiveness regression task, we used an initial learning rate of 1×10^{-4}, a batch size of 32, weight decay of 0.01, and an increased dropout rate in the transformer encoder layers.

For the gender and ethnicity classification tasks, slightly different hyperparameters were adopted to better suit their categorical nature: a batch size of 16, initial learning rate of 3×10^{-4}, dropout rate of 0.3, and weight decay of 1×10^{-4}. Training was extended to a maximum of 250 epochs, with early stopping and the same learning-rate scheduling strategy.

Across all tasks, regularization was introduced via dropout and data augmentation applied only to the training split, including random horizontal flipping, random rotations, and a custom symmetry-preserving transformation intended to maintain the balance of facial features. Model performance was monitored using task-specific metrics at each epoch, and training was halted early if no improvement was observed.

5.0.2 Evaluation Protocol

To evaluate the proposed predictor model, we followed two standard evaluation protocols provided with the SCUT-FBP5500 dataset:

(1) **60/40 Split Protocol** — 60% of the images are used for training and the remaining 40% are held out for testing. This setting enables direct comparison with previous works that utilize the same split.

(2) **5-fold Cross-Validation Protocol** — the dataset is randomly partitioned into five equally sized folds. In each fold, 80% of the images are used for training and 20% are used for testing. The procedure is repeated five times so that each fold serves as the test set exactly once. The final results are averaged over all five folds to obtain a robust estimate of generalization performance.

For the facial attractiveness prediction (regression) task, three metrics were used: Pearson Correlation (PC), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). For the gender and ethnicity classification tasks, standard classification metrics were computed, including Accuracy, Precision, Recall, and F1-score.

Regression Metrics

The **Mean Absolute Error (MAE)** is defined as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

where y_i is the ground-truth score, $\hat{y}_i$ is the predicted score, and n is the number of samples.

The **Root Mean Squared Error (RMSE)** is given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

which maintains the same units as the target variable.

The **Pearson Correlation (PC)** coefficient measures the linear dependence between predicted and true scores:

$$\mathrm{PC} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{9}$$

where $\bar{y}$ and $\bar{\hat{y}}$ denote the mean of the ground-truth and predicted scores, respectively. PC ranges from -1 to 1.

Classification Metrics

The **Accuracy** is defined as:

$$\mathrm{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{10}$$

where TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

The **Precision** is defined as:

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \tag{11}$$

The **Recall** (or sensitivity) is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{12}$$

The **F1-score**, which represents the harmonic mean of Precision and Recall, is given by:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{13}$$

5.0.3 Facial Attractiveness Task Evaluation

In this section, we report the results of the facial attractiveness prediction task obtained using different Vision Transformer (ViT) variants and input resolutions.

Follow the standard 60/40 train–test protocol on SCUT-FBP5500 (60% train, 40% test). For the best-performing configuration, we additionally report results under the 5-fold cross-validation protocol (Table 5.3). We evaluate using Pearson correlation (PC), mean absolute error (MAE), and root mean squared error (RMSE).

Following preprocessing, all facial images were cropped, aligned, and resized to meet the input resolution requirements of our respective ViT configurations. The pretrained ViT models—originally trained on face-recognition datasets—were fine-tuned for attractiveness regression, allowing the networks to adapt their learned representations to beauty score prediction. Performance results under the 60/40 protocol are summarized in Table 5.2, and the best-performing configuration was further evaluated using the 5-fold protocol (Table 5.3).

Table 5.2: Performance comparison of ViT variants on the facial attractiveness prediction task using the 60/40 train–test protocol. Best values in each column are highlighted in bold.

Model	**Patch**	**PC ↑**	**MAE ↓**	**RMSE ↓**
ViT-Tiny	16	0.9024	0.2422	0.3008
	32	0.9050	0.2230	0.2980
ViT-Small	16	0.9153	0.2150	0.2760
	32	0.9380	0.1865	0.2302
ViT-Base	16	0.9423	0.1759	0.2143
	32	**0.9549**	**0.1601**	**0.2047**

To evaluate the facial attractiveness prediction models, we followed the two standard evaluation

protocols provided with the SCUT-FBP5500 dataset. The dataset contains 5,500 facial images, each annotated with attractiveness scores by an average of 60–70 human raters.

For the 60/40 train–test protocol, 60% of the images are used for training and 40% are held out for testing. A small portion of the training set is reserved as a validation set for hyperparameter tuning and early stopping.

For the 5-fold cross-validation protocol, the dataset is divided into five equally sized folds. In each iteration, four folds (80% of the images) are used for training—with a small validation split—and the remaining fold (20%) is used exclusively for testing. This process is repeated five times so that each fold serves as the test set exactly once, and the final reported results are averaged over all folds.

Within the training portion of each fold, a further split was performed to create a validation set, which was used to monitor training progress, tune hyperparameters, and apply early stopping to prevent overfitting. This ensured that validation and test results were computed on distinct, unseen data. The 5-fold cross-validation framework not only provided a more reliable estimation of generalization performance but also enabled fair comparisons between different Vision Transformer (ViT) configurations, as all models were trained and evaluated on identical fold partitions.

Table 5.3: 5-fold cross-validation results for the **best-performing model** (ViT-Base with Patch size of 32, and 224×224 resolution) on the SCUT-FBP5500 dataset. The model was pretrained on the VGGFace2 dataset. Metrics are reported for each fold along with their average.

Fold	**PC** ↑	**MAE** ↓	**RMSE** ↓
Fold 1	0.9540	0.1545	0.2044
Fold 2	0.9599	0.1627	0.2138
Fold 3	0.9492	0.1650	0.2005
Fold 4	0.9510	0.1582	0.2097
Fold 5	0.9588	0.1610	0.2054
Average	**0.9545**	**0.1602**	**0.2067**

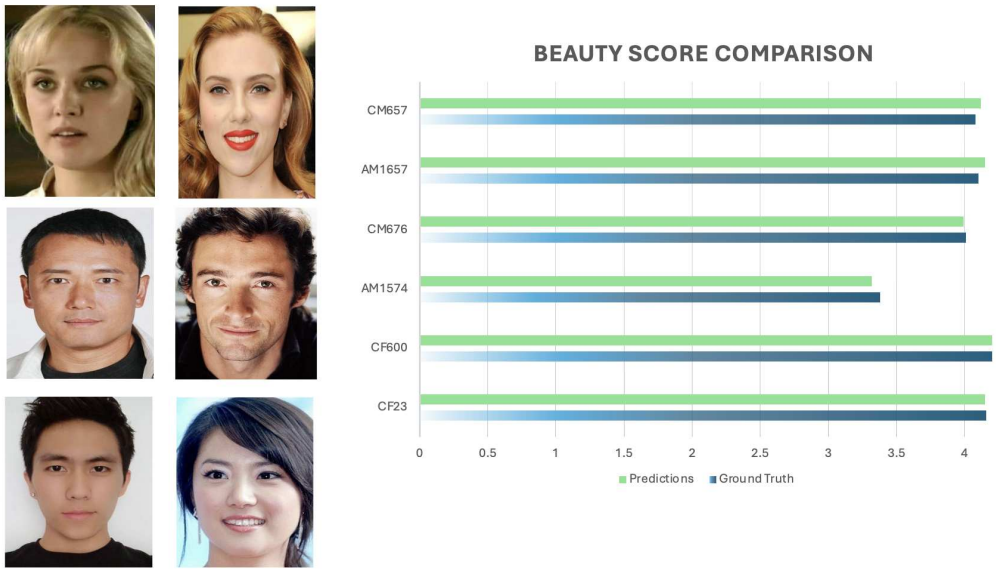

Figure 5.1: Visual comparison between ground-truth attractiveness scores and model predictions on sample subjects from the SCUT-FBP5500 dataset.

5.0.4 Multi-task Learning Evaluation

The multi-task learning framework is designed to jointly perform three tasks: facial attractiveness prediction, gender recognition, and ethnicity identification. The primary task, facial attractiveness prediction, is formulated as a regression problem, while gender and ethnicity recognition are treated as classification problems. The motivation for this setup is that jointly learning these related facial attributes can encourage the model to extract richer and more generalizable facial representations, improving the robustness of attractiveness prediction.

The shared feature extractor is based on the best-performing model from the single-task experiments, ViT-Base, with a patch size of 32, with 224×224 resolution, fine-tuned on the SCUT-FBP5500 dataset. The output of the final transformer encoder block was passed to three task-specific output heads:

- A regression head for attractiveness prediction.
- A classification head for gender recognition (two classes: male, female).
- A classification head for ethnicity identification (four categories).

Each head consists of fully connected layers with dropout regularization.

For attractiveness prediction, the **Mean Squared Error (MSE)** loss was used:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{14}$$

where y_i is the ground-truth attractiveness score, $\hat{y}_i$ is the prediction, and n is the number of samples.

For gender and ethnicity recognition, **Binary Cross-Entropy (BCE) Loss** was applied:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{15}$$

where $y_i \in \{0, 1\}$ represents the ground truth binary label and $\hat{y}_i \in (0, 1)$ denotes the predicted probability.

The total multi-task loss is computed as the weighted sum of the individual task losses:

$$\mathcal{L}_{\text{total}} = \lambda_B \mathcal{L}_{\text{MSE}} + \lambda_G \mathcal{L}_{\text{BCE,gender}} + \lambda_E \mathcal{L}_{\text{BCE,ethnicity}}, \tag{16}$$

where λ_B, λ_G, and λ_E are hyperparameters controlling the contribution of each task. A higher weight was assigned to attractiveness prediction to emphasize its primary importance.

Since facial attractiveness prediction is the main objective of our framework, a larger weight is allocated to this task compared to the auxiliary classification tasks. Through empirical tuning, we observed that the best trade-off between all three tasks was achieved when setting $\lambda_B = 0.6$, $\lambda_G = 0.2$, and $\lambda_E = 0.2$. This configuration yielded the lowest MAE and RMSE values for attractiveness prediction while maintaining strong classification performance for gender and ethnicity. These optimal weights are adopted for all subsequent multi-task experiments.

Table 5.4: Comparison of single-task and multi-task learning configurations for beauty, gender, and ethnicity prediction. Metrics for beauty prediction are reported using PC, MAE, and RMSE. Best results are highlighted in bold.

Task	λ_B	λ_G	λ_E	**PC** ↑	**MAE** ↓	**RMSE** ↓
Beauty (baseline)	1.0	0.0	0.0	0.9545	0.1602	0.2067
Beauty, Gender, Ethnicity	0.8	0.1	0.1	0.9558	0.1601	0.2061
	0.6	**0.2**	**0.2**	**0.9621**	**0.1599**	**0.2043**
	0.5	0.3	0.2	0.9541	0.1664	0.2130

The results in Table 5.4 demonstrate that introducing auxiliary gender and ethnicity prediction objectives consistently boosts attractiveness regression performance relative to single-task learning. The configuration with $\lambda_B = 0.6$, $\lambda_G = 0.2$, and $\lambda_E = 0.2$ achieves the best balance, yielding the lowest MAE and RMSE along with the highest Pearson correlation (PC). This suggests that a moderate emphasis on the auxiliary tasks encourages the network to learn richer and more discriminative facial representations, which translate into improved attractiveness prediction accuracy.

Table 5.5: 5-fold cross-validation results for gender recognition and ethnicity identification using the multi-task learning framework with ViT-Base Patch32-224.

Task	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**
Gender Recognition (%)	99.82	99.94	99.97	99.88	99.85	**99.89**
Ethnicity Identification (%)	99.89	99.84	99.92	99.82	99.86	**99.86**

5.0.5 Attention Map Visualizations Across Model Layers

In transformer-based computer vision models, understanding where the network focuses its attention is a key step toward interpreting the model's decision-making process. For Vision Transformers (ViTs), attention maps offer a visual representation of the regions in an image that the model deems most relevant during both training and inference. This section examines the attention map visualizations for our top-performing model, used for facial attractiveness prediction, gender classification, and ethnicity classification. By analyzing attention across layers, we gain insights into how the model's focus evolves and how it relates to salient facial features.

The self-attention mechanism in ViTs, originally developed for natural language processing, allows the model to capture global dependencies by attending to relationships between all image patches. Through multiple transformer layers, the model learns to emphasize patches containing features critical for attractiveness prediction, such as eye symmetry, jawline structure, and skin smoothness. Attention map visualization serves two purposes: (1) improving model interpretability, and (2) verifying whether the model focuses on semantically relevant facial areas rather than unrelated background regions.

5.0.6 Understanding Attention Maps

An attention map is generated by extracting the attention weights from a self-attention layer and projecting them back to the spatial domain of the input image. For ViTs, each layer contains multiple attention heads, each focusing on different aspects of the input. Early heads might highlight local patterns such as edges or color gradients, while deeper heads capture high-level semantic cues, such as facial regions associated with attractiveness ratings. Observing these maps allows us to evaluate whether the learned attention aligns with human visual intuition.

5.0.7 Visualization Technique

To visualize the model's focus, we computed the attention weights for the [CLS] token relative to all image patches across all transformer layers. These weights were then reshaped to match the patch grid and upsampled to the full image resolution. A heatmap was overlaid on the original image, where warmer colors indicate stronger attention weights. This method enables a visual interpretation of how the model processes facial information at each stage.

5.0.8 Layer-wise Analysis

The ViT-Base Patch32-224 model consists of twelve transformer layers. By visualizing the attention maps for each layer, we observe the following progression:

- **Early Layers (1–4):** Attention is dispersed across multiple regions of the face and occasionally extends to the background. These layers primarily capture low-level visual cues such as

edges, contours, and basic textures.

- **Middle Layers (5–8):** Attention becomes more concentrated on prominent facial regions, including the eyes, nose, hair, and mouth. The model begins to recognize structural patterns that contribute to attractiveness prediction.
- **Later Layers (9–12):** Attention maps exhibit a refined and highly focused pattern, often highlighting areas such as eye symmetry, jawline, and skin texture. At this stage, the model concentrates on features most relevant to the regression and classification tasks.

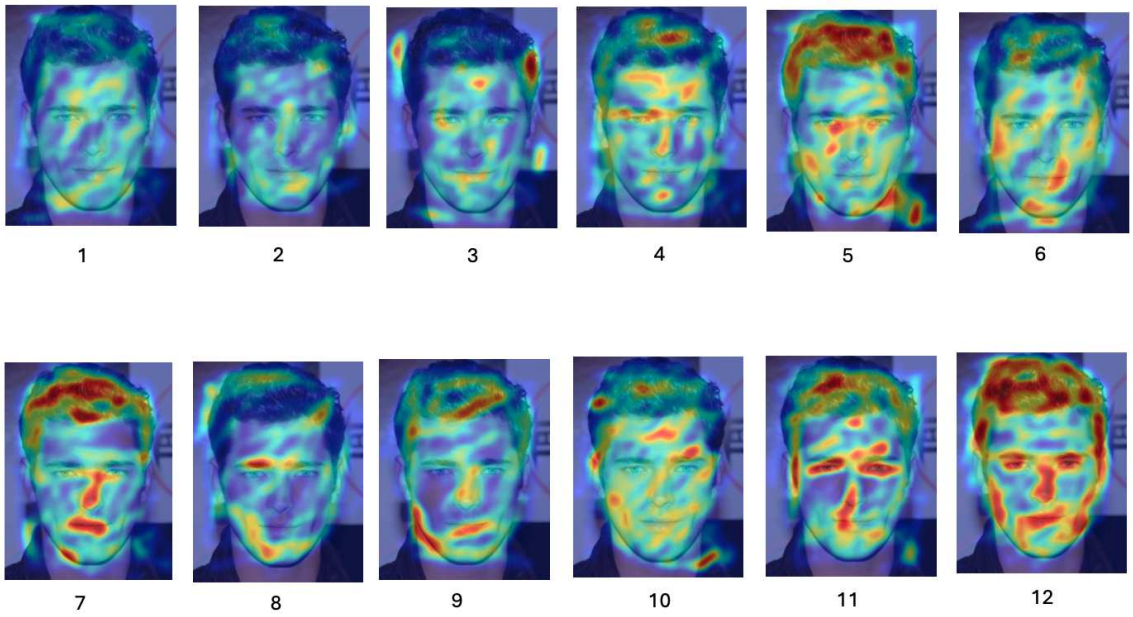

Figure 5.2: Layer-wise attention maps from the ViT-Base Patch32-224 model, showing the transition from dispersed to highly focused attention across early, middle, and later layers.

Figure 5.2 presents attention maps for all twelve layers of the ViT-Base Patch32-224 model on an example image from the SCUT-FBP5500 dataset. The evolution of attention—from a broad, exploratory focus in early layers to concentrated, task-relevant focus in later layers—provides an intuitive narrative of the model's learning process. This analysis confirms that the model predominantly attends to facial features aligned with established aesthetic principles, supporting the validity of its predictions.

5.1 Comparisons with State-of-the-art Methods

To assess the effectiveness of our proposed Vision Transformer-based multi-task framework, we compared its performance with several recent state-of-the-art methods on the SCUT-FBP5500 dataset.

5.1.1 60%-40% Data Split Evaluation

In the first evaluation protocol, we followed the widely used 60%-40% data split strategy, where 60% of the images (3,300) were used for training and the remaining 40% (2,200) for testing. Table 5.6 presents the results in terms of Pearson Correlation (PC), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). For multi-task models, we also report the accuracy of gender (AccG) and ethnicity (AccE) recognition.

Table 5.6: Performance comparison with state-of-the-art works on SCUT-FBP5500 in terms of PC, MAE, RMSE with a 60%-40% data split.

Method	**PC** ↑	**MAE** ↓	**RMSE** ↓
AlexNet Liang et al. (2018)	0.8298	0.2938	0.3819
ResNet-18 Liang et al. (2018)	0.8513	0.2818	0.3703
ResNeXt-50 Liang et al. (2018)	0.8777	0.2518	0.3325
HMTNet L. Xu et al. (2019)	0.8783	0.2501	0.3263
SEResNeXt50 L. Xu and Xiang (2020)	0.9117	0.2126	0.2813
Vahdati et al. Vahdati and Suen (2020)	0.9392	0.1794	0.2356
Ours (ViT-B32-224 - single task)	**0.9549**	**0.1601**	**0.2047**

5.1.2 5-Fold Cross-Validation Evaluation

We also evaluated our method under the **5-fold cross-validation protocol**, where the dataset is evenly split into five folds, with four folds used for training and one for testing in each iteration. The final results are averaged over the five folds (Table 5.7).

Our approach achieved state-of-the-art performance, with a PC of **0.9545**, MAE of **0.1602**, and RMSE of **0.2067**.

Table 5.7: Performance comparison with state-of-the-art works on SCUT-FBP5500 in terms of PC, MAE, RMSE using 5-fold cross-validation.

Method	**PC ↑**	**MAE ↓**	**RMSE ↓**
Geometric features + SVR Liang et al. (2018)	0.6668	0.3898	0.5132
AlexNet (Liang et al., 2018)	0.8634	0.2651	0.3481
ResNet-18 (Liang et al., 2018)	0.8900	0.2419	0.3166
ResNeXt-50 (Liang et al., 2018)	0.8997	0.2291	0.3017
PI-CNN (Lin et al., 2022; J. Xu et al., 2017)	0.8978	0.2267	0.3016
CNN + LDL (Y.-Y. Fan et al., 2018)	0.9031	0.2201	0.2940
HMTNet (L. Xu et al., 2019; L. Xu & Xiang, 2020)	0.8912	0.2380	0.3141
AaNet (Lin et al., 2019)	0.9055	0.2236	0.2954
MobileNetV2 (Shi et al., 2019)	0.9260	0.2020	0.2660
ResNeXt-50 based R3CNN (Lin et al., 2022)	0.9142	0.2120	0.2800
SEResNeXt50 + ComboLoss (L. Xu & Xiang, 2020)	0.9199	0.2050	0.2704
ResNet-50 based multi-task*	0.9175	0.2086	0.2748
Vahdati et al. (multi-task) (Vahdati & Suen, 2020)	0.9440	0.1742	0.2290
Vahdati et al. (multi-stream + multi-task) (Vahdati & Suen, 2020)	0.9482	0.1670	0.2197
EN-CNN (Boukhari, Chemsa, Taleb-Ahmed, Ajgou, & Bouzaher, 2023)	0.9350	0.1933	0.2482
Diff-FBP (Diffusion Transformer, generative pretrain) Boukhari and chemsa (2025)	0.9220	0.2110	–
Ours (ViT-B-224 - single-task)	**0.9545**	**0.1602**	**0.2067**
Ours (ViT-B-224 - multi-task)	**0.9621**	**0.1599**	**0.2043**

5.1.3 Discussion

The strong performance of our model can be attributed to three main factors:

(1) **Powerful Pre-training** – Our ViT backbone was pre-trained on the Face recognition datasets, which are more relevant to facial analysis than ImageNet, enabling richer facial feature representations.

(2) **Multi-task Learning** – Joint training on attractiveness, gender, and ethnicity encourages the model to learn complementary features, improving generalization.

(3) **Transformer Architecture** – Vision Transformers excel at capturing long-range dependencies and global context, which are crucial for modeling subtle facial attractiveness cues.

Overall, our experiments confirm that Vision Transformers, when properly pre-trained and optimized, can achieve and even surpass the performance of the best CNN-based methods on SCUT-FBP5500, setting a new state-of-the-art benchmark.

Chapter 6

Conclusion and Future Directions

6.1 Conclusion

This thesis presented a Vision Transformer (ViT)-based framework for facial attractiveness prediction, with the SCUT-FBP5500 dataset serving as the primary benchmark. The research addressed the challenge of learning robust attractiveness prediction models from limited data by leveraging transfer learning from large-scale face datasets (VGGFace2) and by applying data augmentation strategies. Multiple ViT variants were evaluated across different configurations of patch size and model scale to analyze their impact on regression performance. The best results were achieved with the ViT-Base Patch32-224 configuration, which demonstrated a strong balance between model capacity, feature representation quality, and computational efficiency.

The attractiveness prediction task was formulated as a regression problem and evaluated using three metrics: Pearson Correlation (PC), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The 5-fold cross-validation protocol confirmed the stability and robustness of the best-performing configuration, with consistent performance across all folds.

In addition to single-task learning, we developed a multi-task learning framework that jointly performs facial attractiveness prediction alongside gender recognition and ethnicity identification. In this setup, the gender and ethnicity tasks act as auxiliary classification problems that guide the shared feature extractor to learn more generalized and attribute-aware facial representations. Using the best ViT configuration for all tasks, the multi-task framework achieved near-perfect accuracy

for gender recognition and ethnicity identification, while maintaining high regression performance for attractiveness prediction. This confirms the effectiveness of multi-task learning in improving feature generalization without compromising the primary task.

Overall, the findings of this thesis highlight the potential of transformer-based architectures for facial attractiveness prediction, especially when combined with auxiliary tasks in a multi-task learning setting. The results also demonstrate the value of transfer learning and careful model selection for achieving state-of-the-art performance on relatively small-scale datasets.

6.2 Future Directions

The results and insights obtained in this research suggest several promising directions for future work:

- **Exploring Lightweight Transformers:** Investigate smaller and more efficient transformer architectures or hybrid CNN-Transformer designs that can offer competitive performance with reduced computational requirements, enabling deployment on resource-constrained devices.
- **Integration of 3D Face Information:** Extend the framework to leverage 3D face representations, which could capture structural attributes such as chin, nose, and forehead shapes more accurately, potentially benefiting applications like aesthetic surgery planning.
- **Analysis of Facial Expressions and Non-Permanent Features:** Study the influence of expressions, makeup, hairstyle, and other transient attributes on attractiveness perception to create models capable of handling a wider variety of facial appearances.
- **Dataset Expansion and Diversity:** Develop a larger and more diverse dataset annotated for attractiveness, gender, and ethnicity, covering a broader range of ages, ethnic backgrounds, and attractiveness levels. This would further improve model generalization.
- **Advanced Data Augmentation:** Explore generative approaches, such as GAN-based augmentation, to balance the dataset and increase the representation of underrepresented attractiveness categories.

- **Natural Language Explanations for Beauty Scores:** Extend explainable AI approaches beyond visual attention maps by generating human-readable, natural language justifications for attractiveness predictions. For instance, instead of only assigning a score of "3" or "4," the system could produce explanations such as "The balanced facial symmetry and smooth skin texture contributed positively, while uneven lighting and partial occlusion reduced the perceived attractiveness." Integrating NLP-based explanation generation would make the model's reasoning accessible to non-experts, improve transparency, and allow stakeholders to better evaluate potential biases or inconsistencies. This direction also opens opportunities to align model explanations with human aesthetic reasoning, bridging computer vision and natural language processing for more interpretable and trustworthy beauty prediction systems.

- **Curriculum Learning Strategies:** Implement curriculum learning approaches, starting with simpler cases (e.g., frontal, neutral expressions) before introducing more complex samples, to improve generalization across all tasks.

By pursuing these directions, future research can build upon the contributions of this work to create more accurate, efficient, and generalizable facial attractiveness prediction systems, while broadening their applicability to diverse real-world scenarios.

6.3 Exploratory Work on 3D Object Reconstruction

In addition to the primary experiments in this thesis, we explored a related task: 3D object reconstruction from 2D images. This task aims to infer the underlying three-dimensional geometry of an object given one or more two-dimensional views, producing a mesh or point cloud representation.

We implemented a preliminary reconstruction framework capable of generating 3D geometric structures from single-view 2D inputs. The detailed methodologies and the specific datasets used in this work will be described in future publications once the study has been fully completed. Figure 6.1 presents anonymized, non-identifiable examples produced by the system, which illustrate the feasibility of extracting spatial shape information from limited visual cues.

Such reconstruction capabilities have potential applications in facial analysis, where 3D shape

descriptors—such as depth, curvature, and volumetric ratios—could complement 2D texture features. Although this work was conducted as a separate project, its integration into the facial attractiveness prediction pipeline represents a promising future research direction.

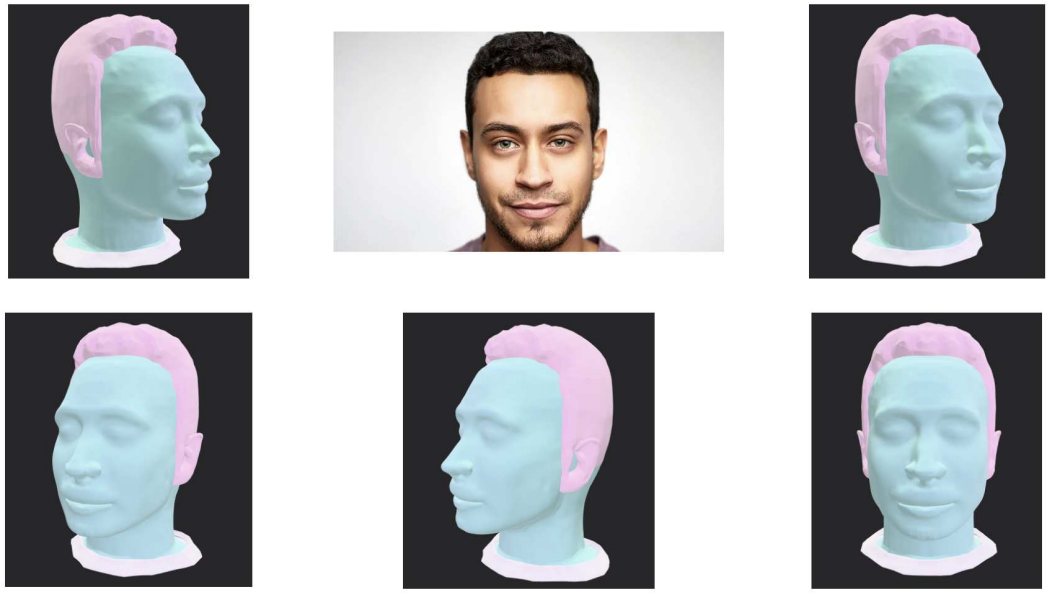

Figure 6.1: A sample of reconstruction model output

References

An, X., Deng, J., Guo, J., Feng, Z., Zhu, X., Jing, Y., & Tongliang, L. (2022). Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *Proceedings of the ieee conference on computer vision and pattern recognition.*

Bainbridge, W., Isola, P., & Oliva, A. (2013, 11). The intrinsic memorability of face photographs. *Journal of experimental psychology. General*, *142*, 1323-1334. doi: 10.1037/a0033872

Bottino, A., & Laurentini, A. (2010, 06). The analysis of facial beauty: An emerging area of research in pattern analysis. In (p. 425-435). doi: 10.1007/978-3-642-13772-3_43

Boukhari, D. E., & chemsa, A. (2025). *Generative pre-training for subjective tasks: A diffusion transformer-based framework for facial beauty prediction.* Retrieved from https://arxiv.org/abs/2507.20363

Boukhari, D. E., Chemsa, A., Taleb-Ahmed, A., Ajgou, R., & Bouzaher, M. t. (2023). Facial beauty prediction using an ensemble of deep convolutional neural networks. *Engineering Proceedings*, *56*(1). Retrieved from https://www.mdpi.com/2673-4591/56/1/125 doi: 10.3390/ASEC2023-15400

Bozkir, M., Karakaş, P., & Oğuz, (2004, 07). Vertical and horizontal neoclassical facial canons in turkish young adults. *Surgical and radiologic anatomy : SRA*, *26*, 212-9. doi: 10.1007/s00276-003-0202-2

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2017). Vggface2: A dataset for recognising faces across pose and age. *CoRR*, *abs/1710.08092*. Retrieved from http://arxiv.org/abs/1710.08092

Dantcheva, A., & Dugelay, J.-L. (2014a, 09). Assessment of female facial beauty based on anthropometric, non-permanent and acquisition characteristics. *Multimedia Tools and Applications*, *74*. doi: 10.1007/s11042-014-2234-5

Dantcheva, A., & Dugelay, J.-L. (2014b, 09). Assessment of female facial beauty based on anthropometric, non-permanent and acquisition characteristics. *Multimedia Tools and Applications*, *74*. doi: 10.1007/s11042-014-2234-5

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, *abs/2010.11929*. Retrieved from https://arxiv.org/abs/2010.11929

Eisenthal, Y., Dror, G., & Ruppin, E. (2006a, 01). Facial attractiveness: Beauty and the machine. *Neural Computation*, *18*(1), 119-142. Retrieved from https://doi.org/10.1162/089976606774841602 doi: 10.1162/089976606774841602

Eisenthal, Y., Dror, G., & Ruppin, E. (2006b, 01). Facial attractiveness: Beauty and the machine. *Neural Computation*, *18*, 119-142. doi: 10.1162/089976606774841602

Eisenthal, Y., Dror, G., & Ruppin, E. (2006c, 01). Facial attractiveness: Beauty and the machine. *Neural Computation*, *18*, 119-142. doi: 10.1162/089976606774841602

Fan, J., Chau, K., Wan, X., Zhai, L., & Lau, E. (2012). Prediction of facial attractiveness from facial proportions. *Pattern Recognition*, *45*(6), 2326-2334. Retrieved from https://www.sciencedirect.com/science/article/pii/S003132031100478X (Brain Decoding) doi: https://doi.org/10.1016/j.patcog.2011.11.024

Fan, Y.-Y., Liu, S., Li, B., Guo, Z., Samal, A., Wan, J., & Li, S. Z. (2018). Label distribution-based facial attractiveness computation by deep residual learning. *IEEE Transactions on Multimedia*, *20*(8), 2196-2208. doi: 10.1109/TMM.2017.2780762

Gan, J., Li, L., Zhai, Y., & Liu, Y. (2014). Deep self-taught learning for facial beauty prediction. *Neurocomputing*, *144*, 295-303. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231214006468 doi: https://doi.org/10.1016/

j.neucom.2014.05.028

Gao, B., Liu, X., Zhou, H., Wu, J., & Geng, X. (2020). Learning expectation of label distribution for facial age and attractiveness estimation. *CoRR*, *abs/2007.01771*. Retrieved from https://arxiv.org/abs/2007.01771

Gao, L., Li, W., Huang, Z., Huang, D., & Wang, Y. (2018). Automatic facial attractiveness prediction by deep multi-task learning. In *2018 24th international conference on pattern recognition (icpr)* (p. 3592-3597). doi: 10.1109/ICPR.2018.8545033

Gunes, H., & Piccardi, M. (2006). Assessing facial beauty through proportion analysis by image processing and supervised learning. *International Journal of Human-Computer Studies*, *64*(12), 1184-1199. Retrieved from https://www.sciencedirect.com/science/article/pii/S1071581906001108 doi: https://doi.org/10.1016/j.ijhcs.2006.07.004

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *CoRR*, *abs/1607.08221*. Retrieved from http://arxiv.org/abs/1607.08221

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, *abs/1512.03385*. Retrieved from http://arxiv.org/abs/1512.03385

Hu, J., Shen, L., & Sun, G. (2017). Squeeze-and-excitation networks. *CoRR*, *abs/1709.01507*. Retrieved from http://arxiv.org/abs/1709.01507

Kagian, A., Dror, G., Leyvand, T., Meilijson, I., Cohen-Or, D., & Ruppin, E. (2008). A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. *Vision Research*, *48*(2), 235-243. Retrieved from https://www.sciencedirect.com/science/article/pii/S0042698907005032 doi: https://doi.org/10.1016/j.visres.2007.11.007

Kao, Y., He, R., & Huang, K. (2016). Visual aesthetic quality assessment with multi-task deep learning. *CoRR*, *abs/1604.04970*. Retrieved from http://arxiv.org/abs/1604.04970

Laurentini, A., & Bottino, A. (2014). Computer analysis of face beauty: A survey. *Computer Vision and Image Understanding*, *125*, 184-199. Retrieved from https://www.sciencedirect.com/science/article/pii/S1077314214000939 doi:

https://doi.org/10.1016/j.cviu.2014.04.006

Liang, L., Lin, L., Jin, L., Xie, D., & Li, M. (2018). SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. *CoRR*, *abs/1801.06345*. Retrieved from http://arxiv.org/abs/1801.06345

Lin, L., Liang, L., & Jin, L. (2022). Regression guided by relative ranking using convolutional neural network ($r^3$3cnn) for facial beauty prediction. *IEEE Transactions on Affective Computing*, *13*(1), 122-134. doi: 10.1109/TAFFC.2019.2933523

Lin, L., Liang, L., Jin, L., & Chen, W. (2019, 7). Attribute-aware convolutional neural networks for facial beauty prediction. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19* (pp. 847–853). International Joint Conferences on Artificial Intelligence Organization. Retrieved from https://doi.org/10.24963/ijcai.2019/119 doi: 10.24963/ijcai.2019/119

Liu, S., Fan, Y., Guo, Z., & Samal, A. (2015, June). 2.5d facial attractiveness computation based on data-driven geometric ratios. In *Proceedings of the international conference on intelligent science and big data engineering* (pp. 564–573). Suzhou, China: Springer.

Liu, S., Fan, Y.-Y., Guo, Z., Samal, A., & Ali, A. (2017a). A landmark-based data-driven approach on 2.5d facial attractiveness computation. *Neurocomputing*, *238*, 168-178. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231217301248 doi: https://doi.org/10.1016/j.neucom.2017.01.050

Liu, S., Fan, Y.-Y., Guo, Z., Samal, A., & Ali, A. (2017b). A landmark-based data-driven approach on 2.5d facial attractiveness computation. *Neurocomputing*, *238*, 168–178.

Liu, S., Fan, Y.-Y., Samal, A., & Guo, Z. (2016, December). Advances in computational facial attractiveness methods. *Multimedia Tools Appl.*, *75*(23), 16633–16663. Retrieved from https://doi.org/10.1007/s11042-016-3830-3 doi: 10.1007/s11042-016-3830-3

Liu, X., Li, T., Peng, H., Ouyang, I. C., Kim, T., & Wang, R. (2019). Understanding beauty via deep facial features. *CoRR*, *abs/1902.05380*. Retrieved from http://arxiv.org/abs/1902.05380

Mahmud, A. (2020). *Query-based summarization using reinforcement learning and transformer*

model (Unpublished doctoral dissertation).

Mao, H., Jin, L., & Du, M. (2009a, 10). Automatic classification of chinese female facial beauty using support vector machine. In (p. 4842-4846). doi: 10.1109/ICSMC.2009.5346057

Mao, H., Jin, L., & Du, M. (2009b, 10). Automatic classification of chinese female facial beauty using support vector machine. In (p. 4842-4846). doi: 10.1109/ICSMC.2009.5346057

Mu, Y. (2013). Computational facial attractiveness prediction by aesthetics-aware features. *Neurocomputing*, *99*, 59-64. Retrieved from https://www.sciencedirect.com/science/article/pii/S092523121200495X doi: https://doi.org/10.1016/j.neucom.2012.06.020

Nezami, K., & Suen, C. Y. (2023). An unbiased artificial referee in beauty contests based on pattern recognition and ai. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100025. Retrieved from https://www.sciencedirect.com/science/article/pii/S2949882123000257 doi: https://doi.org/10.1016/j.chbah.2023.100025

Rizvi, D. Q., & Karawia, A. (2013, 12). Female facial beauty analysis for assesment of facial attractivness..

Rodrigo, M., Cuevas, C., & García, N. (2024). *Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks.* Retrieved from https://openreview.net/forum?id=CCo8ElCT7v

Saari, T.-H., Leppänen, J., Mangs, K., & Savelainen, A. (2008, 03). *Mat-2.4177 seminar on case studies in operations research: Generating aesthetically pleasing lattice structures.* doi: 10.13140/RG.2.2.11427.96808

Schmid, K., Marx, D., & Samal, A. (2008a). Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. *Pattern Recognition*, *41*(8), 2710-2717. Retrieved from https://www.sciencedirect.com/science/article/pii/S0031320307005055 doi: https://doi.org/10.1016/j.patcog.2007.11.022

Schmid, K., Marx, D., & Samal, A. (2008b, 08). Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. *Pattern Recognition*, *41*, 2710-2717. doi: 10.1016/j.patcog.2007.11.022

Shi, S., Gao, F., Meng, X., Xu, X., & Zhu, J. (2019). Improving facial attractiveness prediction

via co-attention learning. In *Icassp 2019 - 2019 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 4045-4049). doi: 10.1109/ICASSP.2019.8683112

Shu Liu, Z. G., Yangyu Fan, & Samal, A. (June 2015). 2.5d facial attractiveness computation based on data-driven geometric ratios. *International Conference on Intelligent Science and Big Data Engineering*.

Sultan, N., Suen, C., Concordia University (Montréal, Q. C. S., & Engineering, S. (2014). *A study on an automatic system for analyzing the facial beauty of young women*. Concordia University. Retrieved from https://books.google.ca/books?id=uM_GzwEACAAJ

Sutic, D., Breskovic, I., Huic, R., & Jukic, I. (2010a, 06). Automatic evaluation of facial attractiveness. In (p. 1339 - 1342).

Sutic, D., Breskovic, I., Huic, R., & Jukic, I. (2010b, 06). Automatic evaluation of facial attractiveness. In (p. 1339 - 1342).

Vahdati, E., & Suen, C. Y. (2019). Female facial beauty analysis using transfer learning and stacking ensemble model. In F. Karray, A. Campilho, & A. Yu (Eds.), *Image analysis and recognition* (pp. 255–268). Cham: Springer International Publishing.

Vahdati, E., & Suen, C. Y. (2020). Facial beauty prediction using transfer and multi-task learning techniques. In Y. Lu, N. Vincent, P. C. Yuen, W.-S. Zheng, F. Cheriet, & C. Y. Suen (Eds.), *Pattern recognition and artificial intelligence* (pp. 441–452). Cham: Springer International Publishing.

Vahdati, E., & Suen, C. Y. (2021). Facial beauty prediction from facial parts using multi-task and multi-stream convolutional neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *35*(12), 2160002. doi: 10.1142/S0218001421600028

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. Retrieved from http://arxiv.org/abs/1706.03762

Xie, D., Liang, L., Jin, L., Xu, J., & Li, M. (2015). SCUT-FBP: A benchmark dataset for facial beauty perception. *CoRR*, *abs/1511.02459*. Retrieved from http://arxiv.org/abs/1511.02459

Xu, J., Jin, L., Liang, L., Feng, Z., & Xie, D. (2015). A new humanlike facial attractiveness

predictor with cascaded fine-tuning deep learning model. *CoRR*, *abs/1511.02465*. Retrieved from http://arxiv.org/abs/1511.02465

Xu, J., Jin, L., Liang, L., Feng, Z., Xie, D., & 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)* (p. 1657-1661). doi: 10.1109/ICASSP.2017.7952438

Xu, L., Fan, H., & Xiang, J. (2019). Hierarchical multi-task network for race, gender and facial attractiveness recognition. In *2019 ieee international conference on image processing (icip)* (p. 3861-3865). doi: 10.1109/ICIP.2019.8803614

Xu, L., & Xiang, J. (2020). Comboloss for facial attractiveness analysis with squeeze-and-excitation networks. *CoRR*, *abs/2010.10721*. Retrieved from https://arxiv.org/abs/2010.10721

Xu, L., Xiang, J., & Yuan, X. (2018a). Crnet: Classification and regression neural network for facial beauty prediction. In R. Hong, W.-H. Cheng, T. Yamasaki, M. Wang, & C.-W. Ngo (Eds.), *Advances in multimedia information processing – pcm 2018* (pp. 661–671). Cham: Springer International Publishing.

Xu, L., Xiang, J., & Yuan, X. (2018b). Transferring rich deep features for facial beauty prediction. *CoRR*, *abs/1803.07253*. Retrieved from http://arxiv.org/abs/1803.07253

Yi, D., Lei, Z., & Li, S. Z. (2015). Age estimation by multi-scale convolutional network. In D. Cremers, I. Reid, H. Saito, & M.-H. Yang (Eds.), *Computer vision – accv 2014* (pp. 144–158). Cham: Springer International Publishing.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2014). Facial landmark detection by deep multi-task learning. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – eccv 2014* (pp. 94–108). Cham: Springer International Publishing.