People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

Department of Electrical engineering

Faculty of Technology

University of Eloued

Thesis to obtaining the LMD doctorate degree in

Electronics

By

Djamel Eddine Boukhari

Thème:

Face image analysis via Transformers: Algorithms and applications for face beauty assessment

Before the jury composed of:

Pr. AJGOU Riadh	*Univ El oued*	*Président*
Pr. CHEMSA Ali	*Univ El oued*	*Supervisor*
Pr. TALEB AHMED Abdelmalik	*Univ Valenncienne France*	*Co- Supervisor*
Pr. OUAFI Abdelkarim	*Univ Biskra*	*Examiner*
Pr. BAARIR Zine-eddine	*Univ Biskra*	*Examiner*
Pr. HETIRI Messoud	*Univ El oued*	*Examiner*

2024

In the Name of Allah

Declaration

Dated:_____ Signature:_____

Dedication

I am feeling great honor and pleasure to dedicate this research work to

My dearest father, Chaabane, and my precious Mom, Sakina

An Inspiration

Out of the huts of history's shame

I rise

I rise

I rise.

Whose endless affection, prayers and wishes have been a great source of comfort for me during my whole education period and my life

Acknowledgments

I would like to thank Pr. Ali Chemsa for having directed my thesis work and for his support during all my years of thesis.

I particularly thank Pr. TALEB-AHMED Abdelmalik, Director of Research at the LAMIH laboratory, Valenciennes, France, for welcoming me into his team, and for his moral support during my thesis work.

I particularly thank Pr. Fadi Dornaika, IKERBASQUE, Basque Foundation for Science, University of the Basque Country UPV/EHU, San Sebastian,Spain, for welcoming me into his team, precious help and for his moral support during my thesis work.

I would like to thank Pr. Ajgou Riadh for his precious help and for agreeing to judge this work.

I would like to thank Pr.Pr. Messaoud Hettiri who has done the honor of chairing this jury.

I would like to thank Pr. Zine Eddine Baarir and Pr. Abdelkrim Ouafi for agreeing to judge this work.

I would also like to thank my colleagues in the CRSTRA and all my friends.

Abstract

The perception of beauty has long been a central topic in human society, shaped by socioeconomic, cultural, and historical influences. Despite the evolving opinions on facial beauty worldwide, understanding the factors behind facial attractiveness remains a key area of interest across disciplines such as psychology, computer science, and evolutionary biology. With advancements in computer vision and deep learning, facial beauty prediction (FBP) has emerged as a cutting-edge field, enabling objective quantification of facial beauty and its underlying factors.

This thesis proposes four novel approaches to facial beauty prediction using deep learning. Two approaches leverage convolutional neural networks (CNNs) integrated with ensemble learning, combining predictions from multiple models to improve accuracy. The remaining two approaches harness the power of Vision Transformers, utilizing attention mechanisms to capture intricate relationships within facial features. Together, these methods enhance feature representation and analysis for robust and reliable facial beauty assessment.

Experiments conducted on the SCUT-FBP5500 benchmark dataset demonstrate the effectiveness of our approaches. We achieve superior performance by comparing various deep learning models, including AlexNet, ResNet-18, and ResNeXt-50. The proposed models yield predictions that closely align with human evaluations, surpassing conventional methods in accuracy and consistency. This thesis underscores the transformative impact of deep learning in facial beauty prediction, offering precise, unbiased, and automated evaluations of facial attractiveness.

keywords : Facial beauty prediction, Convolutional neural networks, Vision Transformers, Ensemble learning, Performance evaluation.

Résumé

La perception de la beauté est depuis longtemps un sujet central dans la société humaine, façonnée par des influences socioéconomiques, culturelles et historiques. Malgré l'évolution des opinions sur la beauté du visage dans le monde entier, la compréhension des facteurs à l'origine de l'attractivité du visage reste un domaine d'intérêt clé dans des disciplines telles que la psychologie, l'informatique et la biologie évolutive. Avec les progrès de la vision par ordinateur et de l'apprentissage profond, la prédiction de la beauté du visage (FBP) est devenue un domaine de pointe, permettant une quantification objective de la beauté du visage et de ses facteurs sous-jacents.

Cette thèse propose quatre nouvelles approches de prédiction de la beauté du visage à l'aide de l'apprentissage profond. Deux approches exploitent les réseaux neuronaux convolutionnels (CNN) intégrés à l'apprentissage d'ensemble, combinant les prédictions de plusieurs modèles pour améliorer la précision. Les deux autres approches exploitent la puissance des transformateurs de vision, utilisant des mécanismes d'attention pour capturer des relations complexes au sein des traits du visage. Ensemble, ces méthodes améliorent la représentation et l'analyse des caractéristiques pour une évaluation robuste et fiable de la beauté du visage.

Les expériences menées sur l'ensemble de données de référence SCUT-FBP5500 démontrent l'efficacité de nos approches. Nous obtenons des performances supérieures en comparant différents modèles d'apprentissage profond, notamment AlexNet, ResNet-18 et ResNeXt-50. Les modèles proposés produisent des prédictions qui correspondent étroitement aux évaluations humaines, surpassant les méthodes conventionnelles en termes de précision et de cohérence. Cette thèse souligne l'impact transformateur de l'apprentissage profond dans la prédiction de la beauté du visage, offrant des évaluations précises, impartiales et automatisées de l'attractivité du visage. **mots-clés :** Prédiction de la beauté du visage, Réseaux de neurones convolutifs, Transformeurs de vision, Apprentissage d'ensemble, Évaluation des performances.

ملخص

لقد كان إدراك الجمال موضوعًا مركزيًا في المجتمع البشري لفترة طويلة، حيث تشكل من خلال التأثيرات الاجتماعية والاقتصادية والثقافية والتاريخية. وعلى الرغم من الآراء المتطورة حول جمال الوجه في جميع أنحاء العالم، فإن فهم العوامل وراء جاذبية الوجه يظل مجالًا رئيسيًا للاهتمام في مختلف التخصصات مثل علم النفس وعلوم الكمبيوتر وعلم الأحياء التطوري. ومع التقدم في مجال الرؤية الحاسوبية والتعلم العميق، برز التنبؤ بجمال الوجه (FBP) كمجال متطور، مما يتيح القياس الموضوعي لجمال الوجه والعوامل الأساسية له. تقترح هذه الأطروحة أربعة مناهج جديدة للتنبؤ بجمال الوجه باستخدام التعلم العميق. يستفيد نهجان من الشبكات العصبية التلافيفية (CNNs) المتكاملة مع التعلم الجماعي، والجمع بين التنبؤات من نماذج متعددة لتحسين الدقة. يستغل النهجان المتبقيان قوة محولات الرؤية، باستخدام آليات الانتباه لالتقاط العلاقات المعقدة داخل ملامح الوجه. تعمل هذه الأساليب معًا على تعزيز تمثيل السمات وتحليلها لتقييم جمال الوجه القوي والموثوق به. تثبت التجارب التي أجريت على مجموعة بيانات SCUT-FBP5500 فعالية مناهجنا. نحقق أداءً متفوقًا من خلال مقارنة نماذج التعلم العميق المختلفة، بما في ذلك AlexNet وResNet-18 وResNeXt-50. تنتج النماذج المقترحة تنبؤات تتوافق بشكل وثيق مع التقييمات البشرية، وتتجاوز الطرق التقليدية في الدقة والاتساق. تؤكد هذه الأطروحة على التأثير التحويلي للتعلم العميق في التنبؤ بجمال الوجه، حيث تقدم تقييمات دقيقة وغير متحيزة وآلية لجاذبية الوجه

Contents

List of Tables

List of Figures

ABBREVIATIONS

FBP Facial Beauty Prediction
CNN Convolutional Neural Networks
ViTs Vision Transformers
NLP Natural Language Processing
MLP The multilayer perceptron
CONV The convolution layer
POOL The pooling layer
FC The fully connected
LOSS The loss layer
ReLU Rectified Linear Unit
RNNs Recurrent neural networks
LSTM Long Short-Term Memory
MSA A Multi-head Self-Attention module (MSA)
FFN feed-forward neural network
SPT Shifted Patch Tokenization
LSA Locality Self Attention
MSMFME Multi-Source Manifold Flexible Manifold Embedding
FSCLDE Feature Selection and Cascaded Deep Discriminant Embedding
NFME Nonlinear Flexible Manifold Embedding
MAE The mean absolute error
RMSE The error root mean square
PC The Pearson correlation

Chapter 1

Introduction

beauty is in the eye of the beholder

– Immanuel Kant: Aesthetics

Facial beauty, a complex and subjective trait, has long fascinated researchers, psychologists, and artists alike. With the advent of advanced technologies in computer vision and artificial intelligence, the exploration of facial beauty prediction methods has become a burgeoning area of research [1]. The ability to quantify and predict perceptions of facial attractiveness holds implications in various domains, including human-computer interaction, healthcare, and the beauty industry. This introduction outlines the significance of facial beauty prediction methods, their current state, and the promising directions for future research [2].

Facial aesthetics play a pivotal role in social interactions, influencing first impressions, interpersonal relationships, and even professional success. The elusive nature of beauty, shaped by cultural, societal, and individual factors, makes it a fascinating and challenging phenomenon to study [3].As we navigate this evolving landscape, the exploration of facial beauty prediction methods not only sheds light on the intricacies of human perception but also challenges us to develop responsible, unbiased, and culturally aware artificial intelligence [4]. This research holds the potential to reshape our understanding of beauty, human-computer interactions, and the ethical considerations that accompany the integration of such technologies into our daily lives [5]. There are two primary obstacles in the examination of facial beauty. Firstly, creating reliable and useful models for assessing beauty is challenging due to the complexity of human perception and the diversity of facial characteristics[6]. Second, many facial reference databases are not suitable for predicting beauty because they are primarily configured for face recognition tasks[7]. As a result, the majority of research on face attractiveness focuses on creating descriptors for facial beauty[8][9]. The perceived attractiveness of a face

is influenced by its symmetry and, to a lesser extent, by sex characteristics [10][11]. Recent advancements in artificial intelligence, particularly in deep learning and computer vision, have paved the way for automated facial beauty prediction. This intersection offers the potential to unravel the intricate patterns and features contributing to perceived beauty [12].

Existing methods leverage machine learning algorithms, particularly deep neural networks, to analyze facial features and patterns. These models are trained on large datasets annotated with beauty ratings, enabling them to learn complex relationships between facial attributes and perceived attractiveness [13][14]. Deep learning architecture has been shaped by the power and flexibility of these algorithms, particularly in relation to convolutional neural networks (CNNs)[15][16]. These algorithms present a novel approach to the problem of predicting facial beauty and have shown promising results for various computer vision applications, including biometrics, captioning, face recognition, object identification, semantic segmentation, and image classification [17][18]. Deep learning models known as Vision Transformers (ViTs) [19] have recently achieved state-of-the-art results in various computer vision tasks, such as segmentation, object recognition, and image categorization. The Transformer architecture, initially developed for natural language processing (NLP) tasks, forms the basis for ViTs [20]. ViTs, on the other hand, significantly modify the Transformer design to enhance its applicability for computer vision[21]. Despite the progress made, challenges such as algorithmic biases, cultural variations, and privacy concerns pose significant hurdles. Ethical considerations surrounding the societal impact of beauty prediction technologies necessitate careful exploration and responsible development.

Accordingly, the study addressed the following research questions:

- Is facial beauty prediction a relatively new area of research in computer vision?
- What are the main facial beauty prediction problems that are currently being tackled using deep learning techniques?
- Which specific deep learning techniques are commonly used in these areas?
- What are some potential future research directions for facial beauty prediction?

The goal of this thesis is to provide an overview of deep learning techniques, particularly vision transformers and convolutional neural networks, in the field

of facial beauty prediction find applications beyond mere aesthetic assessment. They hold potential in virtual try-on experiences, personalized marketing, and healthcare diagnostics, showcasing the diverse and impactful applications of this technology. The journey of facial beauty prediction is far from complete. Future research should delve into multimodal approaches, incorporating cultural sensitivity, ensuring interpretability, and addressing ethical concerns. Collaboration across disciplines and the integration of user feedback will be instrumental in shaping the next generation of facial beauty prediction methods.

1.1 Research motivation

Facial beauty prediction research focuses on understanding the factors that influence perception of attractiveness in faces. Evolutionary-based research suggests that facial attractiveness affects a range of social outcomes, such as mate choices, hiring decisions, and social exchange. Factors that influence attractiveness judgments include symmetry, averageness, skin color/texture, and cues to personality. Research also indicates that individual differences in facial attractiveness can be influenced by factors such as hormone levels, fertility, personality, visual experience, familiarity, and imprinting [22].
Machine learning methods, such as deep self-taught learning, are used to predict facial beauty by analyzing geometric features of faces. Some studies also explore the use of color cues to predict facial attractiveness [23][24]. The motivation behind this research is to better understand the complex factors that contribute to facial attractiveness and to develop more accurate methods for predicting it [25]. The research motivation behind facial beauty prediction (FBP) stems from various fields and interests, some more ethically concerning than others [26]:

1. Understanding the Perception of Beauty:

    - Cognitive Science and Psychology: Researchers in these fields aim to understand the underlying mechanisms of human beauty perception. By studying FBP models and their outputs, they can gain insights into how humans perceive and evaluate facial features, potentially leading to a better understanding of social judgments and preferences.
    - Cultural Studies and Anthropology: Exploring how beauty standards vary across cultures and investigating their historical and social contexts can be facilitated by analyzing FBP models trained on diverse datasets. This can spark valuable discussions about cultural relativism and the evolution of beauty ideals.

2. Technological Applications:

    - Computer Vision and Machine Learning: FBP research pushes the boundaries of computer vision and machine learning by tackling complex tasks like facial image recognition, landmark detection, and feature extraction. These advancements contribute to broader applications in areas like facial recognition software, image editing tools, and even autonomous driving systems (for pedestrian detection).
    - Entertainment and Social Media: Some companies have expressed interest in FBP technology for applications like personalized recommendations in beauty-related domains (e.g., suggesting makeup styles or hairstyles). However, such applications raise significant ethical concerns.
    - Ethical Concerns: The pursuit of FBP is not without ethical criticisms. Some of the major concerns include:
    - Potential for Bias and Discrimination: FBP models trained on subjective and potentially biased datasets can perpetuate existing societal biases related to beauty standards. This can further disadvantage individuals who don't conform to these narrow standards, leading to discrimination in areas like employment, social interactions, and even access to opportunities.
    - Commodification of Beauty: FBP technology risks further objectifying individuals and reducing them to quantifiable beauty scores. This can have detrimental psychological effects on individuals, especially those struggling with body image issues or societal pressures to conform to unrealistic beauty standards.
    - Privacy and Misuse: The potential misuse of FBP technology in surveillance systems or targeted advertising based on beauty scores raises concerns about privacy and potential social manipulation.

1.2 Problem statement

The use of beauty can be traced back to around 4,000 BC.The problem statement in facial beauty prediction research revolves around the challenges posed by the subjective and complex nature of evaluating attractiveness in faces. Some key points from the search results include:

- Indefinite Evaluation Criterion: Facial beauty analysis faces challenges due to its indefinite evaluation criterion, making it difficult to establish a clear standard for assessing attractiveness.

- Small-Scale Databases: Many existing studies rely on small-scale facial beauty databases, which hinder the effective modeling of structural information for facial beauty prediction.
- Burden of Landmarking and Optimization: A significant number of facial beauty prediction algorithms require burdensome landmarking or expensive optimization procedures, impacting the scalability and practicality of these methods.
- Subjectivity and Complexity: The subjectivity of beauty perception and the complexity of determining attractiveness variables remain poorly understood issues in facial attractiveness research.
- Lack of Consensus on Features: There is a lack of consensus on the most relevant features for predicting facial attractiveness, with conflicting results on the importance of geometric, textural, and holistic facial attributes

Addressing these challenges is crucial for advancing facial beauty prediction research and developing more accurate and reliable methods for assessing attractiveness in faces.

1.3 Contribution of the research

Our contributions are motivated by the complex challenge of predicting facial beauty using deep learning techniques. The integration of deep learning methods, such as ensemble convolutional neural networks and vision transformers, has enhanced the accuracy and efficiency of facial beauty prediction models, leading to more reliable outcomes. We emphasize the importance of facial beauty resources, noting that the absence of extensive facial beauty datasets hinders research progress. The thorough investigation that followed led to the conclusion that deep learning approaches have been well examined and have demonstrated great accuracy in the literature. The research on predicting facial beauty makes significant contributions to the field by addressing key challenges and advancing the accuracy of attractiveness prediction in faces. We have introduced feature-based computer models for facial beauty analysis, which enable quick and effective estimation of facial beauty indices through proposed predictive models. Our research has focused on data-driven facial beauty analysis, including prediction, retrieval, and manipulation techniques to enhance the comprehension and prediction of attractiveness in faces. These contributions collectively enhance the understanding and prediction of facial beauty, paving the way for more sophisticated and accurate methods to assess attractiveness in faces. Motivated by the above, we focus in this thesis on deep learning methods,

such as an ensemble of convolutional neural networks and vision transformers. Other notable contributions include:

Major contributions

1. **The algorithm using an ensemble of CNNs named EN-CNN**
   Our approach involves utilizing four models: DenseNet201, InceptionV3, MobileNetV2, and EfficientNetB7. According to the SCUT-FBP5500 benchmark dataset, the results achieved by this new approach in terms of the Pearson coefficient are superior to those obtained by CNN methods. This reveals that the suggested EN-CNNs model can be successfully applied in a variety of face recognition applications.

2. **Facial Beauty Prediction Using an Ensemble of Deep Convolutional Neural Networks (E-CNN)**
   The innovative approach involves using deep learning techniques to predict facial beauty. The primary focus of the research on predicting facial beauty using deep convolutional neural networks is to create an ensemble regression model for estimating facial beauty. The study aims to investigate the effectiveness of transfer learning techniques in predicting facial beauty and to combine the predicted scores of networks with a three-branch network (InceptionV3, MobileNetV2, and S-CNN) trained with loss functions. The goal is to optimize hyperparameters for pre-trained models to classify facial beauty and to create an ensemble (E-CNN) that can predict scores in facial beauty more accurately than previous baseline approaches. The research aims to enhance the congruence of beauty assessment with human judgment and improve the performance of facial beauty prediction models.

3. **Facial Beauty Prediction Based on Vision Transformer**
   Vision transformers play a crucial role in image classification due to their capacity to model long-range dependencies, deliver high performance, provide innovative topology, demonstrate adaptability, and achieve state-of-the-art results compared to traditional methods. We propose to apply vision transformers for facial beauty prediction. These strengths make vision transformers a compelling choice for researchers and practitioners seeking to enhance the accuracy and efficiency of facial beauty prediction models. VIT-FBP is a promising new approach to predicting facial beauty. It is more accurate and interpretable than previous methods, and it has the potential to be used in a variety of applications.

4. **Vision Transformers with Small-Size Datasets for Facial Beauty**

Prediction
STP-LSA-ViT-FBP is a deep learning model that utilizes Vision Transformers with a small dataset to predict facial beauty. It is based on the Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA) which are suggested in this research. They successfully address the issue of the lack of locality inductive bias and allow the system to learn entirely from scratch even on a small-size dataset. STP-LSA-ViT-FBP works by first using spatially shifted input images, where the original input image is an illustration of four diagonal shifts. These SPTs are then fed into a transformer, which learns to encode the features into a representation that captures the local structure of the face, as well as the context in which it is viewed. Finally, the transformer output is fed into a linear layer to predict the facial beauty score. STP-LSA-ViT-FBP has been shown to outperform state-of-the-art methods on several facial beauty prediction benchmarks. It is also more interpretable than previous methods because the transformer allows for the disentanglement of local features.

Minor contributions

- **Comprehensive overview of the state of the art**.We discuss the concept of facial beauty and deep learning techniques, with a focus on deep neural networks and vision transformers. In particular, we describe techniques that have been proposed to extract and compare features that emerge from different beautiful faces.

1.4 Thesis Organization

Our thesis is organized according to the following chapters:

- The first chapter introduces the research topic, provides the background of the study and research motivation, presents the problem statement, outlines the research objectives and questions, discusses the contribution of the research, and explains the organization of the thesis.
- The second chapter of the thesis discusses the state-of-the-art facial beauty prediction methods based on deep learning, focusing specifically on convolutional neural networks and providing a detailed presentation of vision Transformers.
- The third chapter discusses the methodology used in this research. The proposed approaches for optimizing facial beauty prediction methods.

- In the fourth chapter, the data collection process and the experimental setup are discussed along with the results obtained from the experiments.
- The fifth chapter discusses the limitations of the research, presents conclusions, and outlines future work.

1.5 Publications

The research presented here resulted in three conference papers and two papers submitted to a journal.

1.5.1 International conferences

1. Djamel Eddine Boukhari, Ali Chemsa and Zine Eddine Baarir, "Self-Supervised Facial Beauty Prediction Using Nearest-Neighbor Contrastive Learning", IEEE ECTE-Tech24, 17-18 December 2024, ALGERIA.
2. Djamel Eddine Boukhari, Ali Chemsa and Zine Eddine Baarir, "MobileViT architecture for Facial Beauty Prediction", IEEE ICTIS 2024, 14-15 December 2024, ALGERIA.
3. Djamel Eddine Boukhari, Ali Chemsa and Zine Eddine Baarir, "Fusion Vision Transformers and Convolutional Neural Networks for Facial Beauty Predictions", MDPI The 5th International Electronic Conference on Applied Sciences, 4-6 Dec 2024.
4. Djamel Eddine Boukhari, Ali Chemsa and Zine Eddine Baarir, "Semi-supervised facial beauty prediction using contrastive pretraining with SimCLR", MDPI The 5th International Electronic Conference on Applied Sciences, 4-6 Dec 2024.
5. D. E. Boukhari, A. Chemsa, A.Taleb-Ahmed, R.Ajgou, and M. T.Bouzaher.Facial Beauty Prediction Using an Ensemble of Deep Convolutional Neural Networks. Engineering Proceedings, 56(1), 125, 2023.

1.5.2 Journal paper

1. Djamel Eddine Boukhari, et al. "An Ensemble of Deep Convolutional Neural Networks Models for Facial Beauty Prediction." Journal of Advanced Computational Intelligence and Intelligent Informatics 27.6 (2023): 1209-1215.

2. D. E. Boukhari, A. Chemsa, A.Taleb-Ahmed, R.Ajgou, and M. T.Bouzaher.Facial Beauty Prediction Using an Ensemble of Deep Convolutional Neural Networks. Engineering Proceedings, 56(1), 125, 2023.

3. Djamel Eddine Boukhari, Ali Chemsa, and Riadh Ajgou. "Facial Beauty Prediction Based on Vision Transformer."International Journal of Electrical and Electronic Engineering and Telecommunications Vol. 13, No. 3, 2024

Chapter 2

State of the art of facial beauty prediction methods

2.1 Introduction

The human face plays a crucial role in our everyday lives, where an inherent drive for beauty, particularly facial beauty, shapes human behavior. As demand for cosmetic procedures has surged in recent years, a deeper understanding of beauty has become essential in medical fields. Facial attractiveness, or facial beauty prediction (FBP), has emerged as a research area with numerous potential applications. However, it remains a significant challenge in computer vision, largely due to the limited availability of public databases for FBP and the small scale of existing experimental datasets. Additionally, the assessment of facial beauty is subjective, as personal preferences vary widely among individuals [1]. Deep learning methods have demonstrated exceptional abilities in feature extraction and representation, though most prior studies have concentrated on analyzing isolated aspects of facial aesthetics, with limited comparisons between different techniques. The remaining sections of Chapter 2 are organised as follows: Section 2 provides a brief history of research on attractiveness and difficulties in Section 3, Section 4 explains the methodology used in selecting facial beauty prediction methods and Section 5 ends Chapter 2.

2.2 History of Face Beauty Research

For centuries, the nature of beauty has been a subject of debate among philosophers, scientists, and artists. The question of whether beauty is purely subjective or not, famously expressed in Margaret Wolfe Hungerford's oft-cited statement "Beauty is

in the eye of the beholder" (1878), originated in Greece in the 3rd century BCE [27][28][29]. Immanuel Kant's Critique of Judgment (1790) is a seminal work on aesthetics in which he draws a sharp line between reason and feeling. While earlier in his career he felt beauty was purely subjective and gave pleasure, in this work he argues that aesthetic pleasure is something we can expect others to experience as well [30][31]. In declaring an object to be beautiful, we think we have a "reason for demanding a similar delight from everyone," according to Kant [32]. Kant also believed that natural beauty is inherent in a thing, while artistic beauty is a representation of that thing [33] .

In the Roman school, harmony results when a number of diverse parts are unified into a coherent pattern, and beauty arises from this harmony, which is both the one and the many. Life cannot survive without a harmonious pattern of integration in form, so beauty should be found everywhere in the living world. While individuals may have varying opinions on the attractiveness of a particular face, at the population level, there are observable patterns in preferences for facial beauty. The contemporary culture of beautification, including cosmetic and plastic surgery, as well as digital retouching, has significantly influenced our notions of facial beauty [34].

Since the dawn of civilization, humans have been drawn to beauty. The beauty of the human face has inspired poets, painters, and philosophers for millennia. While smooth hair, clean skin, and attractive eyes used to be the essence of beauty, today it includes a wide range of cosmetic procedures, such as hair extensions, face contouring, eyebrow shaping, and eyelash extensions [35] . There are now solutions that can enhance the appearance of any facial feature.

2.2.1 Neoclassical Canons of Facial Attractiveness

The neoclassical canons of facial attractiveness have been extensively documented in art history and aesthetic theory [36]. These canons were first defined by the Greek sculptor Polykleitos during the fifth century BCE and have since been utilized as a benchmark for facial beauty across a variety of cultural contexts [37]. Specifically, the neoclassical canons of facial attractiveness consist of eleven facial proportions that were heavily influenced by Egyptian aesthetics [38]. As depicted in Figure 2.1, these canons include guidelines such as the horizontal line across the eyes that split the head into equal halves, and the positioning of the nose in the center of a face that may be divided into equal thirds [39].

Additionally, the neoclassical canons of facial attractiveness dictate that the head can be divided into equal quarters, with the forehead and nose located in the

middle sections of the head. Moreover, the length of the nose is commensurate with the length of the ear, while the breadth of the nose is equal to the space between the eyes. Furthermore, the width of each eye corresponds to the distance between the eyes, and the width of the mouth is one and a half times the breadth of the nose. Finally, the width of the nose should be one-fourth the width of the face. These canons have had a significant impact on the perception of facial aesthetics and continue to inform contemporary practices in art and design [40].

These neoclassical canons have been widely utilized as a guiding principle for

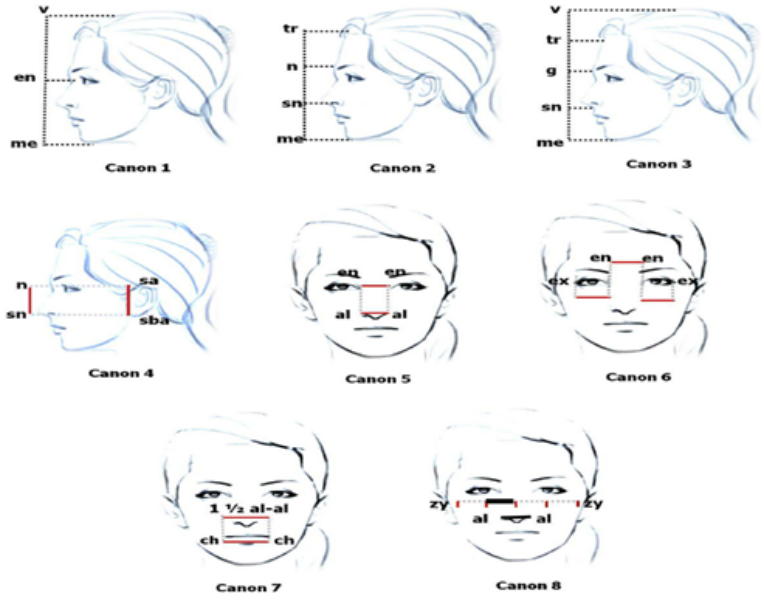


Figure 2.1: Neoclassical canon [41]

artists and scholars alike. Polykleitos, himself, famously employed these canons when creating his celebrated statue Doryphoros [42]. In addition, the concept of facial trisection, sometimes referred to as facial thirds, was developed by the Roman architect Marcus Vitruvius and remains an important tool in the fields of anthropometry and medicine to assess facial aesthetics [43]. According to this concept, a face can be horizontally divided into three equal parts as shown in Figure 2.2. The Renaissance period was marked by a renewed interest in the classical ideals of beauty and proportion. During this time, artists such as Leonardo da Vinci, Leon Battista Alberti, Albrecht Dürer, and others recorded and expanded upon the original Greek canons of proportion [44]. These canons were considered a fundamental aspect of creating realistic and aesthetically pleasing human likenesses in art. Through their work, these Renaissance masters sought to recapture the beauty and harmony of the classical ideals, thus cementing the neoclassical canons of facial proportion as a timeless and essential framework for the portrayal of human form.

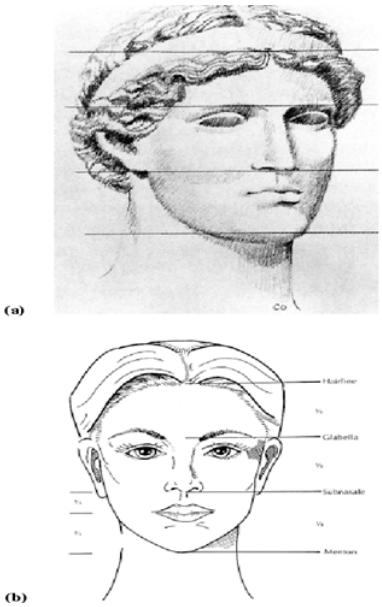


Figure 2.2: a: Facial trisection, as originally described by Vitruvius (c. 70-c. 25 BC) and b: As shown by Powell and Humphries (1984) [45]

2.2.2 The Golden Ratio

The concept of ideal proportions has a rich history dating back to ancient times, with one enduring idea being the significance of the golden ratio in facial beauty [46]. In the context of harmonious relationships between various elements, the term "proportion" is used to compare the size or quantity of distinct sections of an object, as well as to describe the link between them. Mathematically defined as the equality of the form "nine is to three as six is to two," the golden ratio is said to possess aesthetically beautiful harmony [47][48]. This ratio was precisely described by the founder of geometry, Euclid of Alexandria, around 300 BCE as part of a formalized deductive system [49][43].
During the Renaissance, the golden ratio underwent a significant change in direction. No longer was it solely confined to the realm of mathematics [50]. Instead, the ratio found its way into explanations of natural phenomena and into the arts [51]. While there is some debate over whether artists truly used the golden ratio, the neoclassical principles of facial proportions were firmly established using a similar framework. Indeed, the golden ratio ($\phi = 1.618$) was believed to underpin facial beauty, as exemplified by the famous artwork of the Mona Lisa shown in Figure 2.3 [52].

Thus, the history of the golden ratio demonstrates its enduring significance and multifaceted nature, with implications not only in mathematics, but also in the fields of aesthetics and the arts.

Figure 2.3: Mona Lisa and golden ratio (by Leonardo Da Vinci) [53].

2.2.3 Difficulties

As an emerging research field, Facial beauty prediction inevitably suffers significant challenges, encompassing three main categories: classification, regression, and ranking. To advance facial beauty prediction, researchers must overcome several obstacles: Firstly, the field suffers from a lack of resources, with limited literature available on the topic. Existing research studies often focus on specific aspects of facial beauty prediction, with little comparative analysis between different methods. To address this limitation, future research should strive to develop comprehensive models that can accurately predict facial beauty.

Secondly, there is a paucity of public databases suitable for computerized studies of facial beauty. Adequate variability in attractiveness can only be achieved with diverse and extensive datasets. Access to such datasets would enable researchers to develop more robust and generalizable facial beauty prediction models. Thirdly, facial beauty analysis is distinct from other facial analysis tasks, such as face recognition, age estimation, and facial expression recognition. Unlike these tasks, facial beauty models prioritize prediction accuracy, interpretability, and practicality. Consequently, future research must focus on developing facial beauty prediction models that not only deliver accurate predictions but also provide insights into the factors that contribute to facial attractiveness.

In summary, overcoming the challenges associated with facial beauty prediction requires considerable effort due to the field's novelty and lack of resources. However, further research in this area will enable the development of comprehensive and reliable models that have potential applications in multiple domains, including psychology, artificial intelligence, and cosmetic surgery.

2.2.4 Definition of Facial Beauty Prediction

Facial beauty prediction can be defined as a field which involves predicting an individual's facial beauty using various techniques and algorithms. It is a very interesting and useful application as well because it can help both the cosmetic surgery and the psychiatric fields. In the former case, patients desiring improvement in their appearance can get an informed estimation of their facial beauty depending on the suggested surgical interventions, using the developed software; while in the latter, people with dysmorphic conditions, that is, conditions that involve an excessive concern about body image, can also get help by the objective assessment provided by facial beauty prediction [1]. This is important because when the psychiatrist tries to assess the seriousness of the condition and the amount of psychological distress a patient may be experiencing, they need to make a judgement on if the level of concern regarding the physical appearance is realistic. But in reality, he said, there is currently no putative measure which allows an objective assessment of facial beauty; so the specialists had to rely on the patients' descriptions of what they find ugly and usually report low level of psychological distress when their self-perceived ugliness is dismissed as unreal. This illustrates the significance of the new software to both the cosmetic surgery and the psychiatric fields. Furthermore, what I have discussed here is only some of the potential applications of facial beauty prediction. With such a promising future, I would say researches and advancements will definitely keep going in order to expand the field and to perfect the current techniques[3]. Also, one thing about facial beauty prediction should be remembered is that it is not only an objective measurement about someone's facial beauty; cultural, geographical and social differences could lead to variations in the results. For instance, what is considered beautiful in Europe may differ from what is considered beautiful in Asia. This should be carefully taken into account in the usage of this technology especially in the diagnostic and academic terms so as to avoid negligence on the ethical issue of misguidance due to ignorance of cultural variations. So there are still challenges that need to be overcome in the future but the genuine demand for facial beauty prediction is there and is going to continuously inspire the workers in this field. Also, the ongoing investigations in the biomedical and neuroscience, for example, studies on the genetics and the heritability of facial attractiveness, will help to contribute to and broaden the horizon of the facial beauty prediction.

2.3 Facial Beauty Prediction Methods

As it has already been stated, facial beauty prediction has been an active research topic for several years. This section provides an overview of various techniques for creating facial beauty prediction models, along with their benefits and limitations. Facial beauty prediction models can be broadly categorized as either handcrafted features or deep learning models, as shown in Figure 2.4. It is noteworthy that most studies consider facial beauty prediction as a fully supervised task.

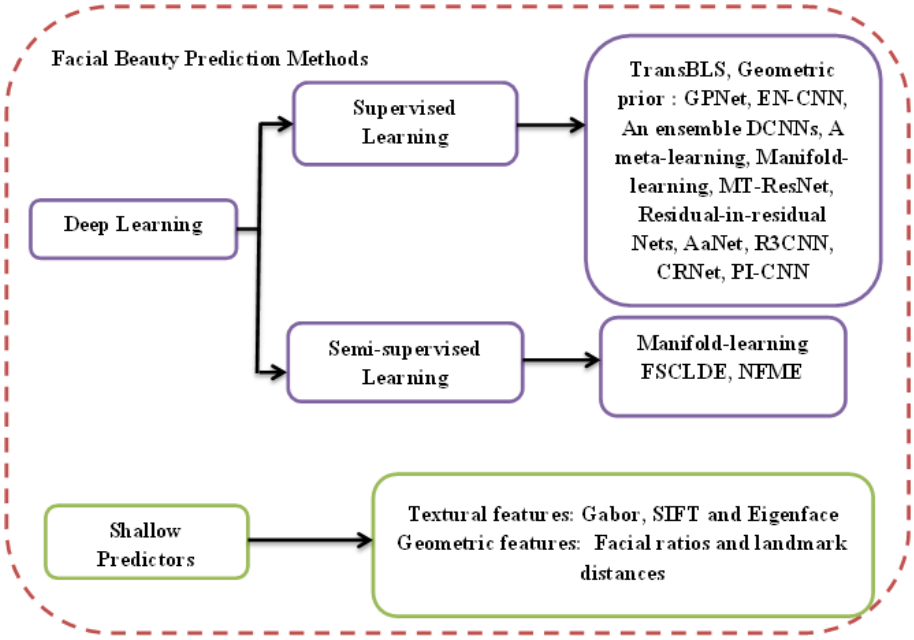


Figure 2.4: Taxonomy of Facial beauty prediction methods

However, while most research studies focus on fully supervised techniques, there have been some recent studies utilizing semi-supervised techniques. Between 2000 and 2010, handcrafted features such as facial ratios and landmark distances were used for facial beauty prediction, while textural features such as Eigenface, SIFT, and Gabor were used between 2005 and 2015. Since 2015, deep learning models have been increasingly utilized for facial beauty prediction using both supervised and semi-supervised techniques, eliminating the need for manually specified facial features. Geometric prior GPNet has been proposed for each deep learning and dataset stream, along with the recently developed Vision Transformer for facial beauty prediction. Their performances have been evaluated over various prevailing benchmarks. Our integrative study highlights notable performance enhancements, and a thorough comparison of current approaches is presented in this paper. To provide context for the current state of research on facial beauty prediction, Figure 2.5 illustrates the number of publications in the field from 2018 to 2023.

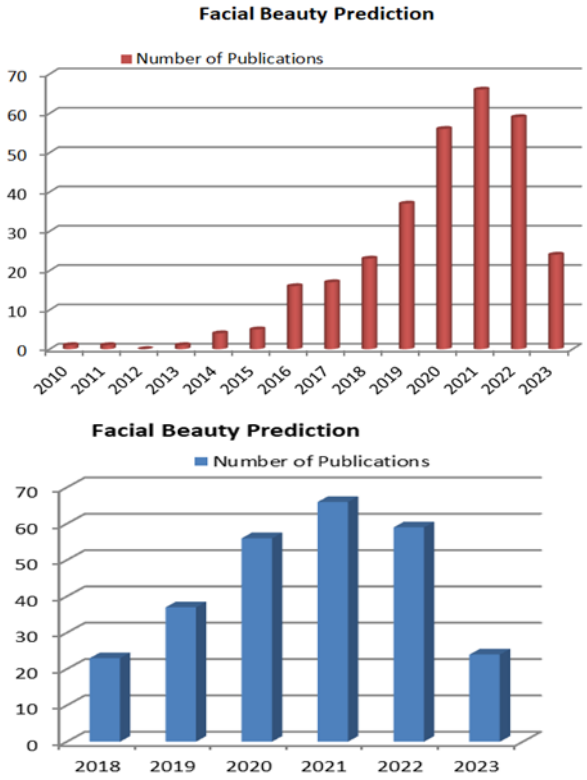


Figure 2.5: Facial beauty prediction publications since 2010 (upper part) and in the last 5 years in Google Scholar (lower part).

The final corpus comprised 265 papers retrieved from Google Scholar. Figure 2.6 compares the number of research studies on facial beauty and facial beauty prediction since the last five years, highlighting the surge in interest in these topics across disciplines such as perception, psychology, biology, artificial intelligence, and more. Facial beauty perception is highly individualized and influenced by social, cultural, and personal factors. Facial beauty prediction is a nascent research topic that requires more investigation and resources.

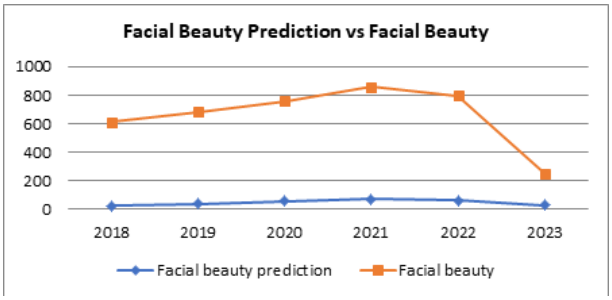


Figure 2.6: Comparison between numbers of research since last five years facial beauty and facial beauty prediction in Google Scholar

Finally, to further contextualize the growth of deep learning and facial beauty research, Figure 2.7 displays Google Trends interests over time for both topics,

revealing a substantial increase in interest since November 2015.

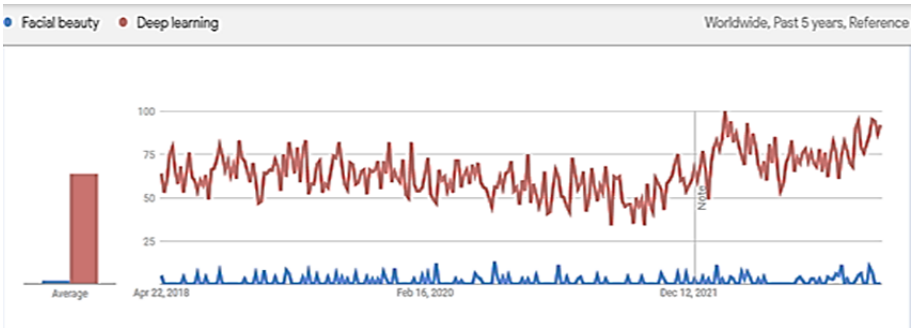


Figure 2.7: Google Trends interest over time for Deep Learning and Face Beauty over the last five years

2.3.1 Deep Learning for Facial beauty prediction

Unlike traditional models that rely heavily on handcrafted features, such as geometric and textural features, deep learning models automatically extract features from input images using deep convolutional neural networks. Numerous studies have demonstrated the effectiveness of deep learning in facial beauty prediction, and a comprehensive comparison of various deep learning-based approaches is presented in Table 2.1. This table summarizes the most important techniques used in these studies, highlighting their respective strengths and limitations, in terms of: Dataset used, Score levels, Techniques used on facial beauty prediction, Validation method (training and test sets) and Metrics. Through this analysis, researchers can gain insights into the most effective techniques for facial beauty prediction and tailor their models accordingly.

Supervised Learning for Facial Beauty Prediction

Facial beauty prediction is a challenging and subjective task that has received considerable attention in recent years. One of the most popular approaches to this task is supervised learning, where the model is trained on a labeled dataset of facial images and their corresponding beauty scores.

1. Pretrained Convolutional Neural Networks (FIAC-Net): This approach leverages the power of pretrained CNNs, specifically networks like AlexNet and VGG16, along with a custom architecture (FIAC-Net). By starting with a pretrained network, the model benefits from extensive prior training on large, general datasets, such as ImageNet, which captures fundamental visual features (e.g., edges, textures). Fine-tuning these pretrained networks on FBP datasets enables the model to adapt these general visual features to

Table 2.1: Compilation of facial beauty prediction approaches developed in the last five years.

Publication	Year	Dataset	Gender	Score levels	Techniques used on facial beauty prediction	Validation method	Metrics
FIAC-Net+ ens. loss[54]	2023	SCUT-FBP SCUT-FBP5500 ME Beauty	F M/F M/F	5 5 10	Refining the performance of three distinct pretrained CNNs, namely AlexNet, VGG16, and FIAC-Net, for the purpose of estimating the beauty score within facial images	Five-fold cross validation	0.9101 Pearson correlation 0.9305 Pearson correlation 0.9260 Pearson correlation
TransBLS [55]	2023	SCUT-FBP5500 LSAFBD	M/F F	5 5	GLAFormer: that combines a transformer with broad learning system to improve the accuracy and reduce overfitting.	Five-fold cross validation.	77.69 % Accuracy 63.97 % Accuracy
GPNet [56]	2023	SCUT-FBP5500	M/F	5	Two branches: a global Swin Transformer branch and a local CNN branch	The split of 60% training and 40% testing Five-fold cross validation	0.9415 Pearson correlation
EN-CNN [57]	2023	SCUT-FBP5500	M/F	5	An ensemble of deep convolutional neural network (CNN) models used the following pre-trained models: DenseNet-201, Inception-v3, MobileNetV2, and EfficientNetB7.	Five-fold cross validation	0.9469 Pearson correlation
Saeed et al.[58]	2023	SCUT-FBP SCUT-FBP5500 ME Beauty	F M/F M/F	5 5 10	An ensemble DCNNs-based regression model	The split of 80% training and 20% testing.	0.879 Pearson correlation 0.886 Pearson correlation 0.888 Pearson correlation
GAN-based ap-proach [59]	2023	SCUT-FBP5500	M/F	5	An innovative GAN-based methodology	–	0.882 Pearson correlation
MSMFME [60]	2023	SCUT-FBP5500	M/F	5	Semi-supervised technique, Flexible manifold embedding, Multi-face beauty similarity	TFive-fold cross validation	91.1 Pearson correlation

Publication	Year	Dataset	Gender	Score levels	Techniques used on facial beauty prediction	Validation method	Metrics
CT Yang et al. [61]	2023	SCUT-FBP5500	M/F	5	The transfer learning, Xception, CNN and attention mechanism models for training	TFive-fold cross validation	DCNN was 0.5295 in RMSE and 18.5% average error in MAPE
Lebedeva et al. [62]	2023	MEbeauty	M/F	10	A meta-learning-based	50, 100, 250 and 500 training samples	50 training was 0.464 Pearson correlation, 100 training 0.574 Pearson correlation, 250 training 0.632 Pearson correlation 500 training 0.685 Pearson correlation
CNN-ER [63]	2022	SCUT-FBP5500	M/F	5	Two branches networkResneXt-50 and Inception-v3	The split of 60% training and 40% testing, 5-fold cross validation.	The split was 0.9207 Pearson correlation 5-fold 0.9250 Pearson correlation
Semi-supervised [64]	2022	SCUT-FBP5500	M/F	5	Two different geometric features and deep face features.	Five-fold cross validation	0.9113 Pearson correlation
ER-BLS [65]	2022	SCUT-FBP5500	M/F	5	Integrating transfer learning and Broad learning system	–	0.9303 Pearson correlation
J Iyer et al.[66]	2021	SCUT-FBP5500	M/F	5	Four regression-based machine learning algorithms are provided with facial ratios as input: Linear Regression, K-Nearest Neighbors, Artificial Neural Network and Random Forest.	The split training 80% and testing 20%	0.7836 Pearson's Correlation, the best for the concatenated features using KNN performs
Wei et al.[67]	2021	SCUT-FBP5500 SCUT-FBP Faces CFD set	M/F F M/F M/F	5 5 - -	features for predicting perceived attractiveness from human portraits, SVM regression and linear regression	Five-fold cross validation	83.72% Beauty score classification 72.20% Beauty score classification 76.17% Beauty score classification 63.48% Beauty score classification

Publication	Year	Dataset	Gender	Score levels	Techniques used on facial beauty prediction	Validation method	Metrics
MT-ResNet [68]	2021	SCUT-FBP5500	M/F	5	A multi-task deep neural network, facial features using machine learning algorithms.	Five-fold cross validation	0.8904 Pearson correlation 98.85% Accuracy of Gender prediction Multi-task 98.18% Accuracy of Gender prediction Single task
FSCLDE [69]	2020	SCUT-FBP5500 SCUT-FBP M2B	M/F F F	5 5 10	Feature Selection Cascade Local Discriminant Embedding, a 1-NN classifier	Five-fold cross validation.	Recognition rate was AF 70.05%, AM 74.60%, CF 70.80% and *CM*72.40% Recognition rate was 74.56% Recognition rate was M2BE 42.58%and M2BW 55.81%
NFME [70]	2020	SCUT-FBP5500 SCUT-FBP M2B	M/F F F	5 5 10	Deep Convolutional Neural Nets, semi-supervised paradigm, Non-linear Flexible Manifold Embedding, 1-Nearest Neighbor (1-NN), Ridge Regression and insensitive Support Vector Regression	Five-fold cross validation. No specific evaluation protocol 2-fold cross-validation	86.72% Pearson correlation 84.64%Pearson correlation a 90%10% 72.90% Pearson correlation a 10%90% 48.05% Both of Pearson correlation, Western 63.22% of Pearson correlation and Eastern 44.55%
Cao et al. [71]	2020	SCUT-FBP5500	M/F	5	Residual-in-residual (RIR) , deeper network , joint attention mechanism, spatial-wise and channel-wise joint	Five-fold cross validation The split of 60% training and 40%testing.	0.9003 Pearson correlation 0.8780 Pearson correlation
Y Zhai et al.[72]	2020	LSAFBD	F	5	local feature fusion and broad learning system, A texture-based local feature fusion method combined with the 2DPCA	–	Accuracy of classification 58.97%
J Gan et al.[73]	2020	LSAFBD	F	5	CNN model, A lighted deep convolution neural network	Training set takes 80% and the other as validation takes 20%.	Accuracy of classification 63.5%

Publication	Year	Dataset	Gender	Score levels	Techniques used on facial beauty prediction	Validation method	Metrics
A-AaNet [74]	2020	SCUT-FBP5500	M/F	5	Pseudo Attribute-aware Convolutional Neural Network, Deep convolution neural networks	Five-fold cross validation.	0.8881 Pearson correlation AaNet + LDL, 0.9055 Pearson correlation ResNet-18 based AaNet 0.9103 Pearson correlation of ResNet-18 based P-AaNet
		SCUT-FBP	F	5			
BeautyNet [75]	2019	LSFBD	F	4	A multiscale CNN architecture, max-feature-map, Transfer learning	The training 90% images from each class, the rest testing.	67.48% classification accuracy, 83.54% Pearson's correlation
R3CNN [76]	2019	SCUT-FBP5500	M/F	5	CNN architecture, Deep Learning Methods	Five-fold cross validation	0.9142 Pearson correlation of ResNeXt-50 based R3CNN
CRNet [77]	2018	SCUT-FBP	M/F	5	Neural network classification and regression network	400 images training and 100 images testing	CRNet achieves a PC with 0.8723 on SCUT-FBP and a PC with 0.482 on HotOrNot dataset
PI-CNN [78]	2017	SCUT-FBP	M/F	5	Deep neural network, a psychologically inspired convolutional neural networks	Five-fold cross validation	0.87 Pearson correlation

more specific beauty-related features, which can include facial symmetry, skin quality, and proportions. These networks excel in capturing hierarchical facial features through convolutional layers, allowing them to distinguish subtle cues in aesthetic assessment. Pretrained models help mitigate the challenge of limited labeled FBP data by transferring knowledge, effectively improving the model's initial performance on beauty-specific tasks. Ensemble methods using multiple CNNs, as seen here, improve robustness and provide a balanced beauty assessment by combining various perspectives from different CNN architectures.

2. Transformers with Broad Learning System (TransBLS): TransBLS combines a transformer-based model (GLAFormer) with a Broad Learning System (BLS). Transformers are designed to handle sequential data by using self-attention mechanisms, which can capture long-range dependencies within facial images. This is particularly useful in beauty prediction as it allows the model to consider relationships across facial features that may be spatially distant. The self-attention mechanism in transformers allows TransBLS to emphasize key aesthetic attributes—like facial symmetry and feature spacing—without losing important contextual information. By adding a Broad Learning System, the model introduces diversity in feature representations, which helps capture different aspects of beauty more comprehensively, ultimately improving robustness and accuracy. The hybrid approach with BLS addresses overfitting, a common issue in FBP with limited data, by enhancing feature diversity and avoiding excessive reliance on any single representation of beauty features. Ensemble of Convolutional Neural Networks (EN-CNN)

3. EN-CNN uses an ensemble of multiple CNN architectures, such as DenseNet-201, Inception-v3, MobileNetV2, and EfficientNetB7. Each of these models has a distinct architectural design and depth, allowing them to capture varied and complementary features within the dataset. Ensemble learning merges their predictions to create a more reliable and stable outcome. Each CNN contributes unique strengths to the ensemble: DenseNet-201 captures dense feature connections, Inception-v3 excels in multi-scale feature extraction, MobileNetV2 is optimized for efficient computations, and EfficientNetB7 offers powerful feature extraction with minimal computational cost. This diversity enhances the model ability to detect a wide range of aesthetic attributes. EN-CNN effectively mitigates the challenge of generalization in FBP by combining models with varied strengths, which reduces overfitting and increases robustness. The ensemble approach also compensates for the limitations of individual CNN architectures, producing a model that can

generalize better across different beauty perceptions.

4. GPNet, introduced by Peng Tianhao et al.[56] (2023), is a hybrid model designed to enhance the accuracy of facial beauty prediction (FBP) through a dual-branch architecture. This innovative model combines the power of Convolutional Neural Networks (CNNs) and Transformers, leveraging the strengths of both local feature extraction and global attention mechanisms. The architecture consists of two main branches:

   Local CNN Branch: This branch focuses on capturing local, fine-grained features from facial images. CNNs are well-suited for this task as they can efficiently learn spatial hierarchies of features. The local CNN branch is responsible for extracting detailed facial features that are crucial for evaluating beauty, such as facial symmetry, texture, and small-scale patterns.

   Global Swin Transformer Branch: In contrast to the local CNN branch, the global branch uses the Swin Transformer, a type of Vision Transformer (ViT) that employs a hierarchical design to capture long-range dependencies and global context in images. The Swin Transformer excels at processing larger-scale, global patterns that could be crucial for understanding overall facial structure and beauty. This global attention mechanism enables the model to consider a broader context when making predictions, which can be particularly important in beauty assessment. Both branches perform multiscale feature fusion, where the outputs from various scales of each branch are combined to form a comprehensive feature representation of the face. This fusion enhances the model's ability to capture both detailed local features and broad, high-level patterns, providing a more accurate prediction of facial beauty. Additionally, geometric regularization is applied to the model to further improve performance. Geometric priors are used to impose structural constraints that help the model better understand the geometric properties of facial features, such as symmetry, proportions, and alignment. This regularization helps refine the feature learning process and contributes to more reliable beauty predictions. Figure 2.8 illustrates the GPNet architecture, showcasing the interaction between the local CNN branch and the global Swin Transformer branch, and highlighting the feature fusion and geometric regularization techniques that make the model effective for facial beauty prediction. This hybrid approach, combining the strengths of CNNs and Transformers, represents a significant advancement in FBP models by leveraging both local detail and global context in the beauty prediction task.

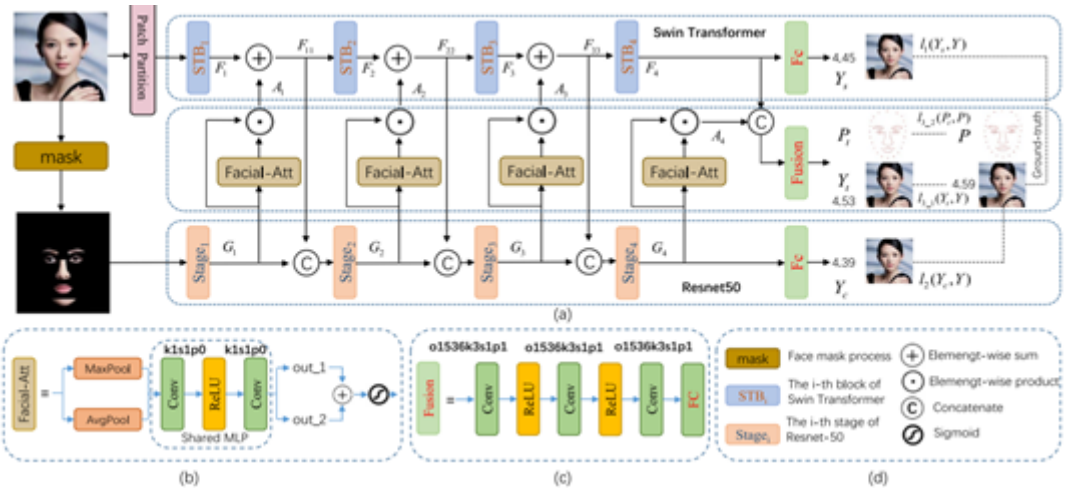


Figure 2.8: GPNet architecture [56].

5. The ensemble DCNNs model, introduced by Saeed et al. (2023) [58], is a regression-based approach for facial beauty prediction that leverages the power of multiple deep convolutional neural networks (DCNNs). The model is designed to improve prediction accuracy by combining the strengths of multiple networks, each contributing unique features to the final decision. The architecture consists of two fine-tuned pre-trained CNNs—AlexNet and VGG16—which have been widely used in computer vision tasks due to their strong ability to extract complex features from images. These models are fine-tuned on the facial beauty prediction task, allowing them to learn the most relevant features for beauty score estimation. In addition to the fine-tuned pre-trained networks, the ensemble model also includes one network built entirely from scratch. This network is trained specifically for the facial beauty prediction task, enabling it to learn features that may not be captured by the pre-trained models. By combining the predictions from these three networks, the ensemble model benefits from the complementary strengths of each network, resulting in more robust and accurate facial beauty predictions. Figure 2.9 in Saeed et al.'s work provides a visual representation of the ensemble DCNNs architecture, showing how the outputs of the individual networks are combined to produce the final beauty score prediction. This approach demonstrates the effectiveness of ensemble learning in improving the performance of FBP tasks by reducing overfitting and capturing a broader range of features from the input images.

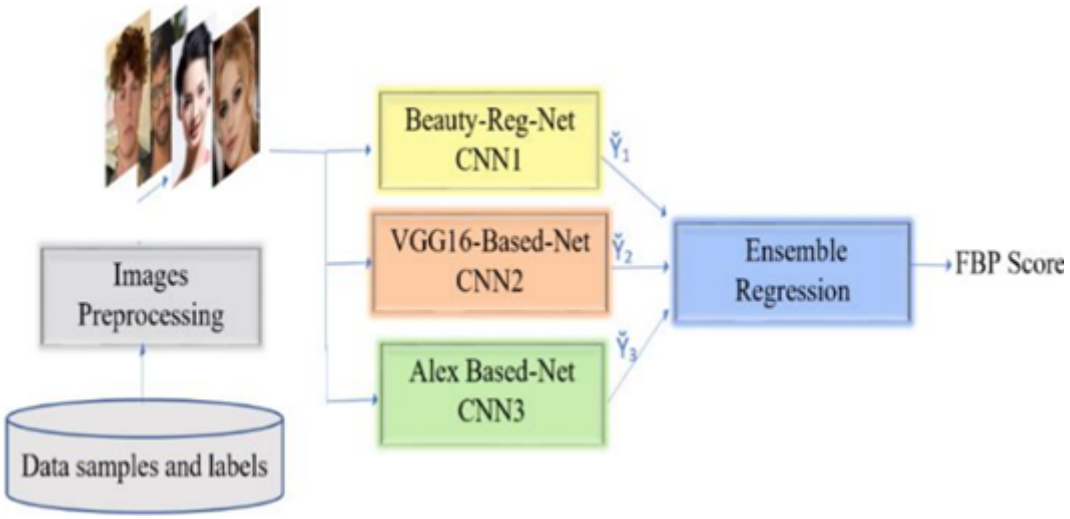


Figure 2.9: An ensemble DCNNs-based regression model [58]

6. The CNN-ER model, introduced by F. Bougourzi et al. (2022)[63], is a dual-branch ensemble CNN framework specifically crafted to enhance accuracy in facial beauty prediction. This model integrates two powerful backbone networks—ResNeXt-50 and Inception-v3—which allow it to leverage diverse feature representations of facial attributes. Each branch in the CNN-ER architecture focuses on capturing different aspects of facial beauty, with ResNeXt-50 specializing in identifying complex facial patterns and Inception-v3 concentrating on multi-scale feature extraction. One distinctive aspect of CNN-ER is its use of four sophisticated loss functions tailored to improve training robustness and precision: A variant of the SmoothL1 loss that reduces sensitivity to outliers and dynamically adjusts its parameters based on training data patterns. This loss function helps the model to be more flexible in managing errors by balancing the penalty for large and small deviations. This loss merges the characteristics of L2 loss for smaller errors and L1 for larger errors, dynamically adjusting during training. It's particularly useful for facial beauty prediction, where subtle errors are more acceptable, but large deviations need greater correction. As a robust loss function, Tukey's biweight loss reduces the impact of extreme outliers, allowing the CNN-ER to focus on capturing accurate beauty estimations for the majority of the data without being skewed by anomalies. A standard loss in regression tasks, MSE penalizes larger deviations between predicted and actual beauty scores, reinforcing the model accuracy. The CNN-ER training approach, employing these four loss functions simultaneously, provides a multi-faceted error evaluation during model optimization. This setup leads to more stable and reliable convergence and achieves a balanced performance across different types of beauty features. The ensemble nature of the CNN-ER, along with

its loss functions, effectively reduces overfitting and ensures that the model generalizes well, making it robust across diverse facial datasets. Figure 2.10 visually outlines how these branches and loss functions are integrated into the CNN-ER framework.

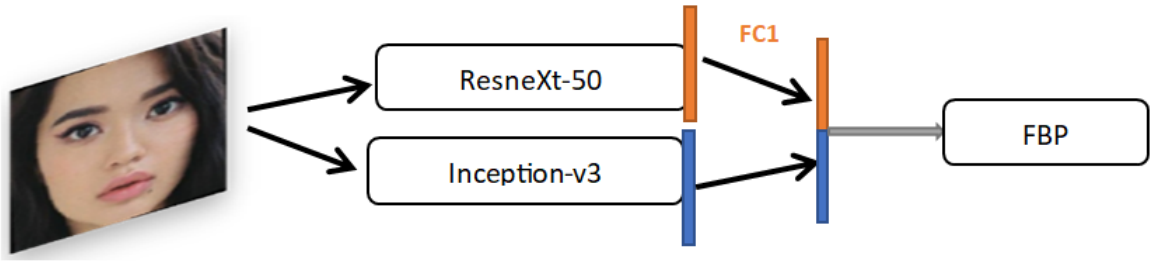


Figure 2.10: the structure of CNN-ER

7. In their study, Cao et al. [71] (2020) introduced the CNN-SCA model, a deeper network architecture designed to enhance facial beauty prediction through a novel residual-in-residual (RIR) structure. This RIR approach stacks multiple residual connections within each module, effectively preserving feature information across layers and mitigating the vanishing gradient problem. The authors utilized the SCUT-FBP5500 dataset, a comprehensive face beauty database with 5,500 facial images labeled with beauty scores, to train and validate their model. To thoroughly assess CNN-SCA effectiveness, two evaluation protocols were applied: a 5-fold cross-validation setup and a train-test split of 80%-20% and 60%-40%. These protocols were employed across three prominent CNN architectures—AlexNet, ResNet-18, and ResNeXt-50—to explore the model's adaptability and performance. A key innovation in CNN-SCA is its combined spatial-wise and channel-wise attention mechanism, which enhances feature extraction by allowing the model to prioritize significant facial attributes in each image. This dual attention approach promotes a deeper comprehension of beauty-relevant features, helping CNN-SCA achieve superior results for facial beauty prediction by focusing on the most informative areas within facial images.

8. The R3CNN model, introduced by Lin et al.[76] (2019), incorporates a relative ranking mechanism into the regression model to enhance facial beauty prediction (FBP) performance. Unlike traditional regression models, which predict an absolute beauty score, R3CNN focuses on relative ranking between images, improving its ability to capture subtle differences in beauty ratings. This method allows the model to learn not only the beauty score of an individual face but also how it ranks relative to other faces, which is crucial

for FBP tasks where small variations can significantly impact predictions. R3CNN is designed to be flexible, allowing it to use existing CNN architectures as its backbone network, making it adaptable to a wide range of face beauty prediction tasks. This adaptability ensures that the model can benefit from the strengths of various pre-trained networks, enabling higher accuracy. The model has shown better performance compared to traditional models on well-known datasets such as SCUT-FBP and SCUT-FBP5500. By integrating relative ranking into the learning process, R3CNN enhances the model's ability to generalize and predict more accurate beauty scores, especially when trained on large, diverse datasets.

Semi-Supervised learning

Semi-supervised learning is a type of machine learning that involves training a model on a small amount of labeled data and a large amount of unlabeled data. This technique has been increasingly applied in the field of facial beauty prediction, where the availability of labeled data is often limited. In this context, F. Dornaika and colleagues have proposed several semi-supervised techniques that have shown promising results.

1. The Multi-Source Manifold Flexible Manifold Embedding (MSMFME) technique, introduced by Dornaika in 2023 [60], is a novel approach for facial beauty prediction (FBP) that leverages manifold learning and graph-based methods. MSMFME aims to enhance the performance of beauty prediction models by incorporating multiple graphs to create a flexible and adaptive manifold embedding model. The model utilizes multiple graphs to represent different sources of information or relationships within the data. By leveraging multiple graphs, the model is able to capture a variety of complex dependencies between facial features, which might not be easily detected using a single graph. These graphs can include different types of relationships such as spatial, texture-based, or even appearance-based features of the face. The core idea of MSMFME is to learn a flexible manifold embedding. Manifold learning aims to map the high-dimensional facial features into a lower-dimensional space, while preserving the intrinsic geometry of the data. This makes the model capable of learning non-linear relationships between facial features and beauty scores. The flexibility aspect refers to the model's ability to adapt its embedding structure based on the available data, which can be both labeled (with beauty scores) and unlabeled (without beauty scores). The MSMFME model can be trained on both labeled and unlabeled data, which allows for semi-supervised learning. This is especially useful in

real-world scenarios where labeled data may be limited, but large amounts of unlabeled data are available. The model can leverage the unlabeled data to improve its feature learning and better predict facial beauty. The model was evaluated using fivefold cross-validation on the SCUT-FBP5500 dataset, which is a widely used dataset for facial beauty prediction containing 5,500 facial images with associated beauty scores. This rigorous evaluation protocol ensures that the model's performance is robust and generalizable across different subsets of the data.

2. The Feature Selection and Cascaded Deep Discriminant Embedding (FSCLDE) technique, proposed by Dornaika et al. in 2020 [69], is a semi-supervised approach designed to improve facial beauty prediction (FBP) by transforming weak and noisy descriptors into more robust and discriminative features. The key innovation behind FSCLDE is its ability to integrate feature selection with deep learning methods to enhance feature quality, which is critical for tasks such as beauty prediction. Cascaded Architecture: FSCLDE employs a cascaded approach where the feature extraction and selection are performed in a sequential manner. This means that after an initial extraction of features, the most relevant and discriminative ones are selected, while irrelevant or noisy features are discarded. This helps in focusing the model's attention on the most informative aspects of the facial features for beauty prediction. Feature Selection: The technique uses feature selection methods to reduce dimensionality and enhance the signal-to-noise ratio. The result is a more focused representation of the face's beauty-related characteristics. Transformation of Linear Approaches: FSCLDE is designed to transform traditional linear techniques into deep variations. This means that the method can take a simpler, linear approach to feature extraction or selection and enhance it with deep learning techniques, which can better capture nonlinear patterns in data, leading to improved performance for facial beauty prediction tasks. Deep Discriminant Embedding: FSCLDE incorporates a deep discriminant embedding technique, which helps the model learn the most discriminative features for facial beauty prediction. This embedding allows the model to project facial features into a new space where the beauty scores are more easily separable, thus enhancing the model's ability to predict beauty accurately. FSCLDE is a semi-supervised method, meaning it can leverage both labeled and unlabeled data. This is especially useful for facial beauty prediction tasks where labeled data may be limited, and the model can make use of large amounts of unlabeled data to improve performance. The method was evaluated on three widely used facial beauty datasets:

SCUT-FBP5500: A large dataset with 5,500 facial images and beauty scores. SCUT-FBP: Another version of the SCUT dataset with facial images and beauty annotations. M2B: A dataset specifically for facial beauty tasks, with images annotated with beauty scores. Evaluation Protocol: FSCLDE was evaluated using a 1-NN classifier (1-Nearest Neighbor classifier), a simple yet effective classification method, with a fivefold cross-validation protocol. This rigorous evaluation protocol ensures that the results are reliable and generalizable across different subsets of data.

3. The Graph-based Nonlinear Flexible Manifold Embedding (NFME) technique, proposed by Dornaika et al. in 2020 [70], is another advanced method designed to improve Facial Beauty Prediction (FBP). This approach builds on the foundation of the Flexible Manifold Embedding (FME) technique, enhancing it with nonlinear methods and graph-based learning strategies, specifically tailored to handle the complexities associated with facial beauty data. Graph Representation: NFME leverages graph-based learning to model the relationships between facial images more effectively. In this context, each facial image can be viewed as a node in a graph, and edges represent similarities or relationships between these images. This graph-based representation helps to capture more intricate patterns and relationships between facial features, crucial for accurate beauty prediction. Nonlinear Learning: Unlike linear techniques, NFME incorporates nonlinear transformations into the manifold learning process. This enables the method to handle complex, high-dimensional data where beauty prediction patterns are not linearly separable, improving the model's ability to capture subtle variations in facial features and beauty scores. The core of NFME is based on the Flexible Manifold Embedding (FME) technique, which is typically used for dimensionality reduction. However, in NFME, this method is modified to handle the nonlinear aspects of the data. The modification is achieved by kernelizing the linear FME, making it more adaptable to the nonlinear relationships present in facial beauty data. By applying kernel functions, NFME can map the input data into higher-dimensional spaces, where nonlinear relationships become more linear and easier to model. This is especially useful when dealing with the complex, high-dimensional features of facial images that linear methods struggle to capture. The technique is specifically designed to work with texture features, which are crucial in facial beauty prediction. Facial beauty often involves complex texture patterns in the skin, eyes, and lips, and NFME's ability to capture these subtle textures leads to more accurate beauty assessments. The use of texture as a primary

feature source allows NFME to focus on important aspects of the face that contribute significantly to perceived beauty, such as smoothness, clarity, and other fine-grained characteristics of facial skin.

Overall, these semi-supervised techniques have shown promising results in facial beauty prediction. Further research is needed to evaluate their effectiveness on larger datasets and in different contexts.

2.3.2 Datasets for Facial Beauty Prediction

In recent years, several datasets have been developed for facial beauty prediction, providing researchers with a valuable resource to evaluate the performance of their models. In this regard, a comparison and analysis of the existing face attractiveness datasets is crucial to determine the most suitable datasets for the task. Table 2.2 summarizes the most important datasets used in numerous studies. The table provides information on the dataset used, the number of raters, the score levels, the facial features used, the validation method (training and test sets), the classification or regression techniques employed, and the metrics used to evaluate the models. These datasets are typically organized by the year they were published and often have limitations regarding the number of face images and the diversity of the individuals represented. The symbol "-" is used to indicate that a dataset contains no information on a specific attribute. In the early stages of facial beauty prediction research, datasets gathered for this purpose contained 100 or fewer face images [83][84][85]. This limited number of images resulted in poor performance and significant bias, especially in deep learning approaches. However, as research in this field progressed, later datasets with different limitations were introduced, containing ten times as many face images. It is worth noting that some datasets are restricted to only one gender and may contain only female images [86][87][88]. Additionally, nearly all facial beauty prediction datasets include ethnicity restrictions. For instance, Zhai [89][90] compiled a dataset of over 20,000 faces, yet they all exhibit an Asian appearance. Currently, the standard SCUT-FBP5500 dataset [79] has gained popularity in facial beauty prediction research. This dataset comprises 5,500 frontal face images at a resolution of 350 x 350 with various attributes, including race (Asian or Caucasian), gender (female or male), and age (15–60). Overall, the use of high-quality, diverse datasets is critical in advancing the field of facial beauty prediction.

As shown in Figure 2.11, the beauty scores for each face in the dataset are the average ratings provided by 60 assessors, using a scale from 1 to 5. This allows for the application of various computational models and diverse approaches to facial

Table 2.2: Summary of facial beauty datasets.

Dataset	Year	# images	Gender	Ethnic	Expression	Age	Rater	Score	Facial features	Validation method	Technique	Metrics
Aarabi et al.[83]	2001	80	female	–	–	–	12	1-4	8 ratios landmarks used in the beauty	40 training set, 40 test set	Modified KNN	91% accuracy of classification images where features have been accurately identified
Gunes et al.[84]	2004	215	female	–	–	All	48	1-10	13 ratios landmarks used in the beauty	10-fold cross validation	Three classifiers: Decision Trees, Multi-Layer Perceptron and Kernel Density Estimators	Accuracy defined in Standardized Distance 0.8102 classifier
Eisenthal et al.[85]	2006	184	female	1 : 92 Cau. American 2 : 92 Cau. Israeli	neutral	Young	1: 28 2: 18	1-7	1: 37 geometric feature + indicators for symmetry 2: The eigenfaces	Leaven-out, with n = 1 for KNN and linear regression and n = 5 for SVM.	Classification using KNN and SVM	correct classifications of 75%to 85% of the images and hybrid predictor 0.65 a correlation with the human ratings
Whitehill et al.[86]	2008	2000	Female/male	80% of the individuals white 10% Asian/5% African/Blacks	46 component movements	18	8	1-4	Gabor features and no geometric feature	5-fold cross validation	SVM regression	The mean Pearson correlation was 0.28 using Gabor features

Dataset	Year	# images	Gender	Ethnic	Expression	Age	Rater	Score	Facial features	Validation method	Technique	Metrics
Gray et al. [87]	2010	2056	Female	Caucasian	All	18-40	30	-3-3	Multiscale single and two layer local filters, eigenfaces	1028 for training, 1028 for testing	Regression model	The correlation ratings: 0.458 for multiscale model
Multi-Modality Beauty (M2B) [88]	2012	1240	Female	Eastern and Western	All	–	40	1-10	Local binary patterns (LBP), Gabor filter	2-fold cross validation	1-NN classifier, Ridge Regression, Neural Network, linear regression	Use the MAE to evaluate the accuracy of the attractiveness prediction
SCUT-FBP [82]	2015	500	Female	Asian	Neutral	–	75	1-5	Geometrical feature + Gabor features	10-fold cross-validation	SVR. Gaussian regression, CNN-based deep learning	PC for traditional machine learning and deep learning is 0.6482 and 0.8187
LSFBD [89]	2016	20000	Female/ male	Asian	All	All	200	0-4	LBP, Eigenfaces and CRBM	5-fold cross validation	Traditional machine-learning method such as KNN or SVM as the preprocessing algorithm	The result of CRBM reach 51.62% (for female) and 52.86% (for male)

Dataset	Year	# images	Gender	Ethnic	Expression	Age	Rater	Score	Facial features	Validation method	Technique	Metrics
SCUT-FBP5500 [79]	2018	5500	Female/male	Asian and Caucasian	Neutral	15-60	60	1-5	18-dimensional ratio feature, Geometric Feature with Shallow Predictor and 40 Gabor feature maps	5-fold cross validation, The split of 60%training and 40% testing.	CNN models with different structures for FBP, AlexNet, ResNet18and ResNeXt-50	The best Pearson correlation ResNeXt-50 0.8997 for 5-fold cross validation and 0.8777 The split of 60% training and 40% testing.
LSAFBD [90]	2020	20000	Female/male	Asian	All	All	200	1-5	LBP and LPQ, Raw Pixel features, HOG, Garbor, SIFT	36,000 training is composed of the training set in DataSet 20000 and its augmented data. 2,000 testing	CNN model	The Pearson correlation coefficient from 0.8594 to 0.8829
ShadowFace3D [91]	2021	6000	Female/male	Asian	–	18-45	20	1-5	3D geometry and 2D texture	Split the 6,000 3D faces into a training (4,200 + (300 + 300)), a validation (400 + (100 + 100)) and a test (400 + (100 + 100))	Deep 3D facial attractiveness prediction , three constituent modules: a 3DFacePoint-Net++ module, a ResNet module and a fusion module	0.849 Pearson correlation coefficient

Dataset	Year	# images	Gender	Ethnic	Expression	Age	Rater	Score	Facial features	Validation method	Technique	Metrics
MEBeauty [92]	2021	2550	Female/male	Caucasian, Asian, Black, Indian, Hispanic and Mideastern	All	All	300	1-10	No manually specified facial features are required	80% training and 20% testing, 90% is used for learning and 10% is used for validation	Deep beauty pattern learning, Different CNN architectures are exploited	0.748 Pearson correlation

The symbol "-" indicates that the associated work contains no information on the specific attribute.

attractiveness prediction. The SCUT-FBP5500 dataset is divided into four subsets: 2000 Asian females (AF), 2000 Asian males (AM), 750 Caucasian females (CF), and 750 Caucasian males (CM), categorized by both race and gender [93][94]. As illustrated in Figure 2.12, the study uses 86 contour points on the face to create a matrix of geometric features. These facial landmarks take into account different facial features and ideal proportions that are relevant to both Asian and Caucasian populations [95]. It is important to note that previous geometric methods did not directly use facial traits as features for predicting attractiveness [96][97]. Figure 2.13 presents facial images from commonly used beauty databases.

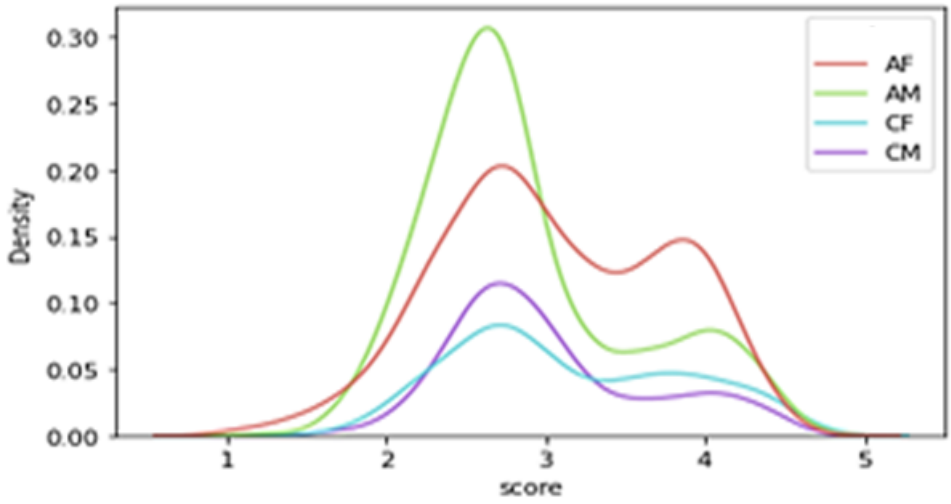


Figure 2.11: The distribution of SCUT-FBP5500 benchmark dataset.

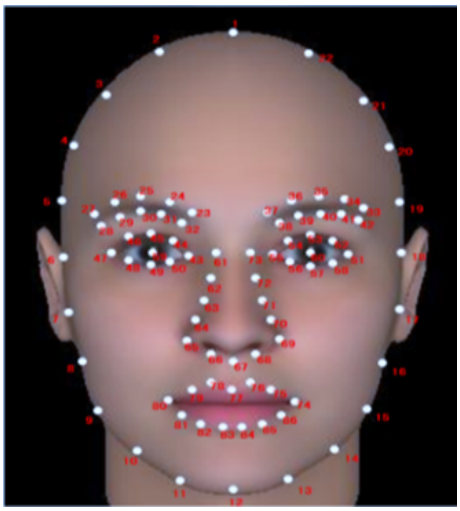
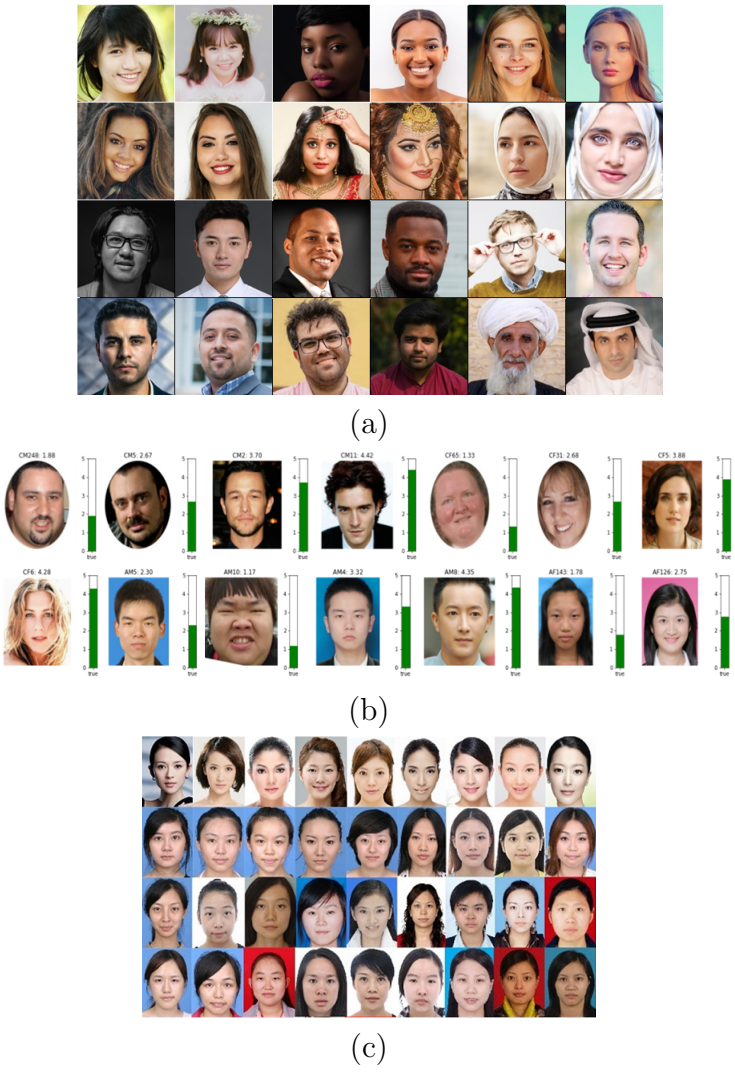


Figure 2.12: The 86 facial points detected in a face image [67] .



(a)

(b)

(c)

Figure 2.13: Image samples of faces from most databases that are used facial beauty prediction, a: MEBeauty dataset, b: SCUT-FBP5500 dataset and c: SCUT-FBP dataset

2.3.3 Overview of popular CNN models for facial beauty prediction

Most facial beauty prediction models employ transfer learning, where pre-trained CNN models are fine-tuned on facial beauty datasets. Typically, CNN models are trained end-to-end to complete classification or regression tasks. However, it is possible to extract deep features by removing the fully connected layers and keeping only the convolutional ones [98][99]. A CNN typically consists of convolutional layers, pooling layers, and fully connected layers. Convolutional layers are the core building blocks of a CNN [100]. It is more common to use models of CNNs already trained and to re-adapt them for the classification problem or other, this is called transfer learning. It involves transferring learning from a model dealing with one problem to another type of problem. This can be done in two ways:

- The extraction of variables from the CNN: here, the CNN is used as an extractor, i.e. a vector is extracted from a certain layer of the model without modifying anything in its structure or its weight and the previously extracted vector is used for a new task.

- Fine-tuning the CNN model: here, the new network is initialized with the weights and structure of the pre-trained model to be used. The structure of the pre-trained model is slightly modified for the new task, and finally the new model is trained for the new task.

Definition and History of Convolutional Neural Networks

1. Definition

   Convolutional neural networks are a type of artificial neural network widely used in the fields of image and video analysis and object recognition. This type of neural network is designed to process data that is typically two-dimensional and uses many different layers to extract different profiles of images and determine and classify their content [101]. Convolutional Neural Networks (CNN) are an extension of MLPs to effectively address the main shortcomings of MLPs. They are designed to automatically extract features from input images, are invariant to slight image distortions, and implement the notion of weight sharing to significantly reduce the number of network parameters. This weight sharing also makes it possible to strongly take into account the local correlations contained in an image [102].The weights are forced to be equal to detect lines, points or corners at all possible places in the image, effectively implementing the idea of weight sharing [103].

2. History of Convolutional Neural Networks The first convolutional neural network was developed in 1998 by French researcher Yann LeCun. This network called LeNet has made it possible to achieve very good performance in character recognition. Although this approach gives results, its progress and evolution have been limited by technological progress in terms of micro-processors, computing power and the lack of accessibility to data in order to be able to train the neurons, however, some researchers continued to work on this model for about two decades [104]. And, with the help of developments in technology but above all with the ever-increasing availability of data, have been able to improve this technique. It was not until 2012 that deep learning was revived by successfully winning the image recognition competition founded by the University of Stamford (Large Squale Visual Recognition Challenge: ILSVRC) thanks to developments in technology and

the ever-increasing availability of data [105]. A new deep learning algorithm explodes records It is a convolutional neural network called AlexNet, largely inspired by the LeNet network. ImageNet comprising 15,000,000 natural images comprising different objects and various scenes (vehicles, animals, etc.) [106]. Today, convolutional neural networks are still the most efficient models for image classification. Google, Microsoft, Facebook, Baidu (the Chinese search engine), Alibaba (Chinese merchant site), Nvidia (graphic processor giant)... use CNNs in their applications [107].

3. Architecture of CNN Convolutional neural networks (CNNs) are built on the principles of multilayer perceptrons (MLPs) and are inspired by the functioning of the visual cortex in vertebrates. While MLPs are effective for image processing, they struggle with large images due to the exponential increase in the number of connections as the image size grows [108]. A convolutional neural network is composed of multiple layers, as illustrated in Figure 2.14.

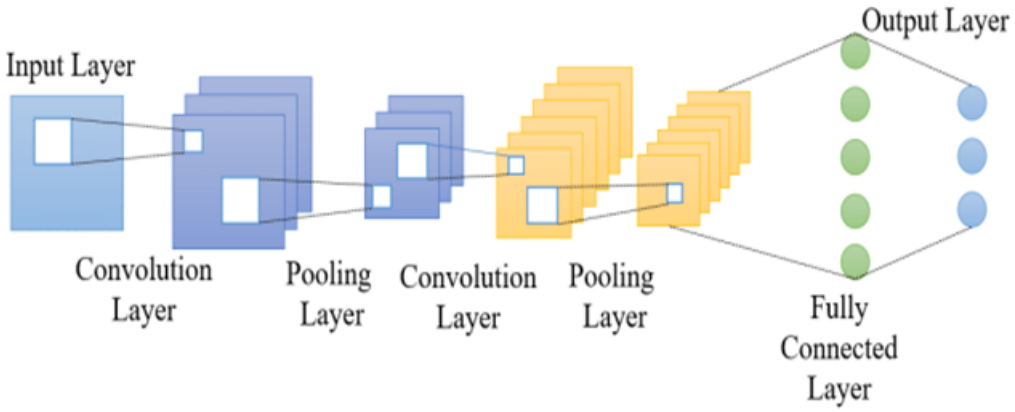


Figure 2.14: Architecture of a convolutional neural network

The architecture of CNN layers is structured by stacking independent processing layers, as described below [109]:

- Convolution Layer (CONV): This layer processes the input data from a receptive field, extracting features through convolutions.
- Pooling Layer (POOL): This layer reduces the size of the intermediate image, typically through subsampling, which helps compress the information.
- Activation Layer (ReLU): Often referred to as ReLU, this layer applies a rectified linear unit (ReLU) activation function to introduce non-linearity, enabling the network to learn more complex features.

- Fully Connected (FC) Layer: This layer is similar to a perceptron and connects all neurons from the previous layer to every neuron in the current layer.
- Loss Layer (LOSS): The loss layer computes the error or loss between the predicted output and the ground truth, which is then used to update the model's weights during training.

4. Benefits of CNNs

   - No Need for Human Supervision: CNNs can automatically learn important features from the data without requiring manual intervention or feature engineering.
   - High Accuracy in Image Recognition and Classification: CNNs excel in tasks like image recognition and classification due to their ability to learn spatial hierarchies of features.
   - Weight Sharing: A significant advantage of CNNs is weight sharing, which reduces the number of parameters, making the model more efficient and preventing overfitting.
   - Minimized Computation: CNNs reduce computational requirements compared to regular neural networks, thanks to techniques like local receptive fields and weight sharing, which helps process large images more efficiently.

   There are a large number of pre-trained CNN networks, the best known are:

5. LeNet: LeNet [104], a 7-layer convolutional network introduced by LeCun et al. in 1998, was initially designed for handwritten digit classification. It is renowned for its ability to recognize numbers, particularly used by banks for processing checks scanned at 32x32 pixel resolution. While LeNet was groundbreaking in its time, the model's ability to handle higher-resolution images is constrained by the need for more convolutional layers and increased computational resources. As a result, its performance is limited when dealing with large or complex image data (as illustrated in Figure 2.15).

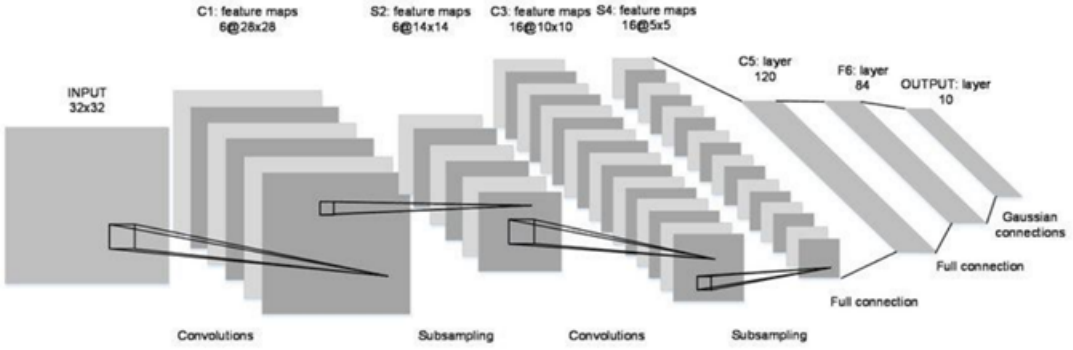


Figure 2.15: LeNet Architecture

6. AlexNet: AlexNet [80] is a deep convolutional neural network designed to handle the complexity of learning difficult objects and their hierarchies. The architecture consists of five convolutional layers, two fully connected hidden layers, and one fully connected output layer. Figure 2.16 provides a visualization of the AlexNet architecture. A key characteristic of AlexNet is its sensitivity to the architecture's structure—removing even a single convolutional layer results in a significant degradation in performance. This emphasizes the importance of each layer in extracting and processing complex visual features.

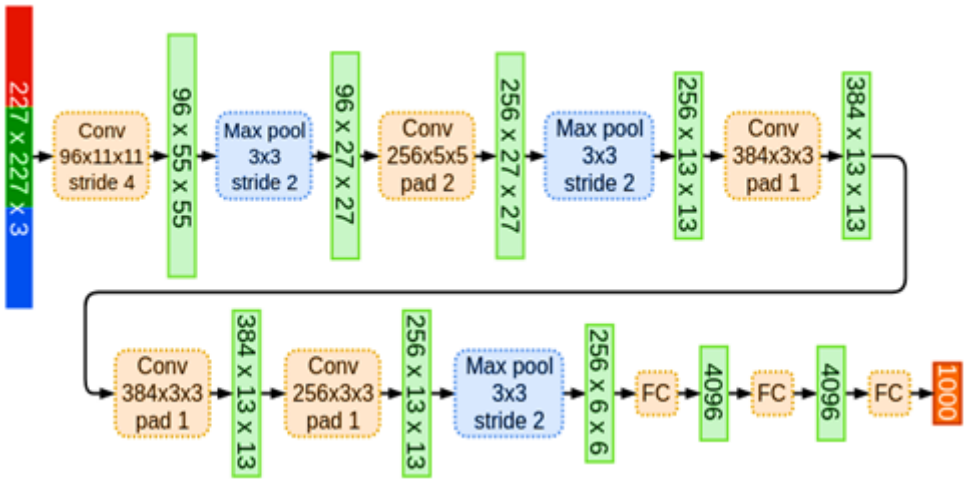


Figure 2.16: AlexNet structure

7. VGGNet: VGGNet [105], introduced by Simonyan and Zisserman in 2014, is characterized by its use of multiple 3x3 convolutional layers stacked in sequence. This design mimics the effect of larger receptive fields (such as 5x5 and 7x7) while keeping the model more efficient and manageable in terms of computational complexity. VGGNet is known for its relatively simple architecture but a large number of parameters, which means it requires

considerable computational power for training. Figure 2.17 illustrates the VGGNet architecture, which is widely recognized for its deep and effective feature extraction capabilities in various image recognition tasks.

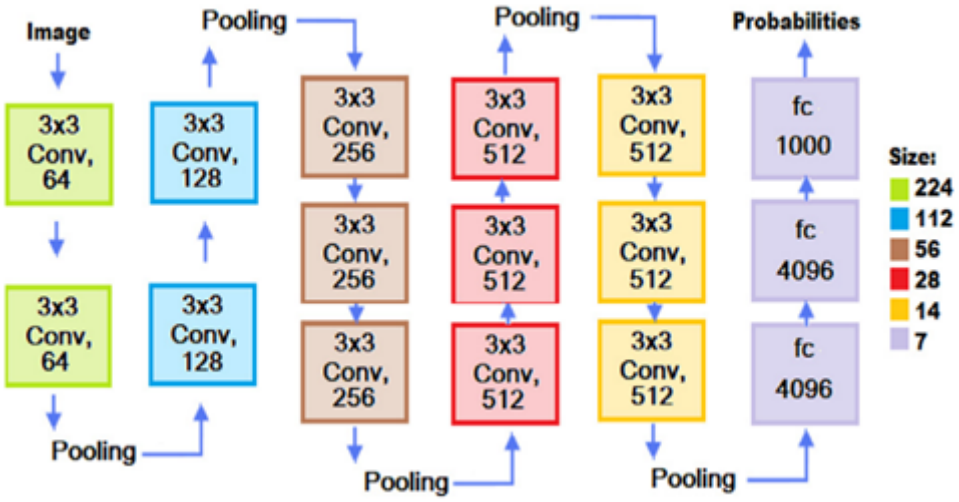


Figure 2.17: VGG-16 layer structure

8. In 2014, Szegedy et al. introduced GoogleNet, a highly efficient deep neural network that incorporated the inception module. This module combines multiple convolutional filter sizes, specifically 1x1, 3x3, and 5x5, along with a pooling layer. The inception module's design helps reduce the number of features and operations at each layer, thus improving computational efficiency and reducing time and cost. GoogleNet, with its 22 layers, addressed the challenge of computational efficiency by optimizing the architecture to minimize the amount of compute required for training and inference. Google has since released several iterations of the Inception architecture, each improving upon the previous version. Inceptionv1 (GoogLeNet), for instance, consisted of 27 layers. Later versions, such as Inceptionv2, Inceptionv3, and Inceptionv4, tackled specific challenges like batch normalization, factorization, and grid size control to further enhance the performance and efficiency of the model [107][110].

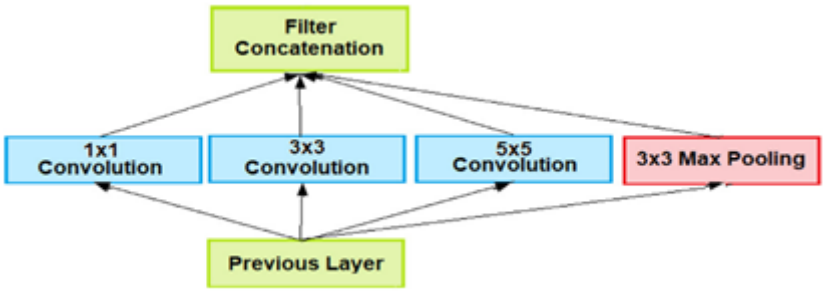


Figure 2.18: Inception module

9. ResNet (Residual Networks), introduced by He et al. in 2015, is a popular neural network architecture, particularly in the context of facial beauty prediction [111]. One of the main challenges in training deep neural networks with many layers is the vanishing gradient problem. As the depth of the network increases, gradients can become very small, making it difficult for the network to learn effectively, which leads to performance saturation or even degradation [112]. ResNet addressed this issue by introducing identity shortcut connections. These connections allow the input to bypass certain layers and be added directly to the output, creating a residual mapping. This innovation helps mitigate the vanishing gradient problem, as the gradient can flow directly through these shortcut connections, making it easier to train much deeper networks. This approach has made ResNet one of the most widely used architectures in deep learning tasks, including facial beauty prediction.

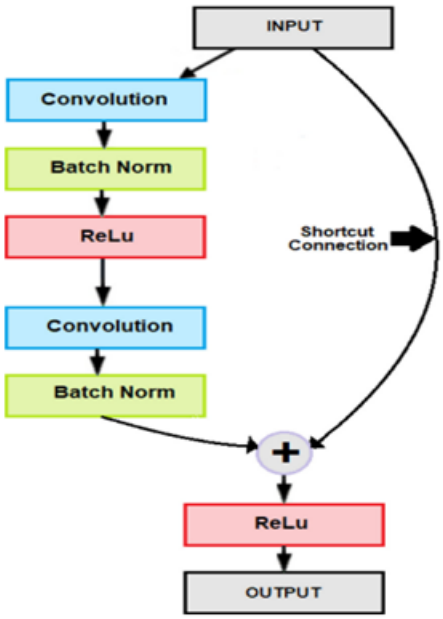


Figure 2.19: Residual learning: a structure block.

ResNet has inspired the development of several modified architectures, one of which is ResNeXt, proposed by Xie et al. [111]. ResNeXt is based on the split-transform-merge technique, similar to the inception modules used in other networks. Unlike the original ResNet, where the outputs from various paths are added together through identity shortcuts, ResNeXt concatenates the outputs of different paths. Furthermore, in ResNeXt, all the paths share the same topology, which differs from ResNet's approach of applying identical operations along the shortcut connections. This architectural modification allows ResNeXt to achieve better performance with fewer parameters while maintaining computational efficiency [113][114].

2.3.4 Vision Transformers

An innovative deep learning model architecture called a Transformer is the first presented in the paper "Attention is all you need". Considering that the Transformer debuted in 2017 [115]. The models developed using this architecture has been published in several articles. For example, the first Vision Transformer from An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [19], BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding [116], and Swin Transformer: Hierarchical Vision Transformer using Shifted Windows [117]..etc. Transformers are deep learning models that are built on the self-attention mechanism.

Motivation

At the moment, deep learning is seen to offer enormous potential for the future as a step towards artificial general intelligence and transformers is one such example. Due to their architecture, they can learn extremely complicated representations and are more generalizable, less prone to overfitting. Recurrent neural networks (RNNs) are already rendered obsolete in natural language models by the Transformer design. Additionally, the Vision Transformer (ViT) has demonstrated superior performance in image classification tasks compared to several Convolutional Neural Networks (CNNs). Figure 3.1 below demonstrates the rise in interest in research on vision transformers.

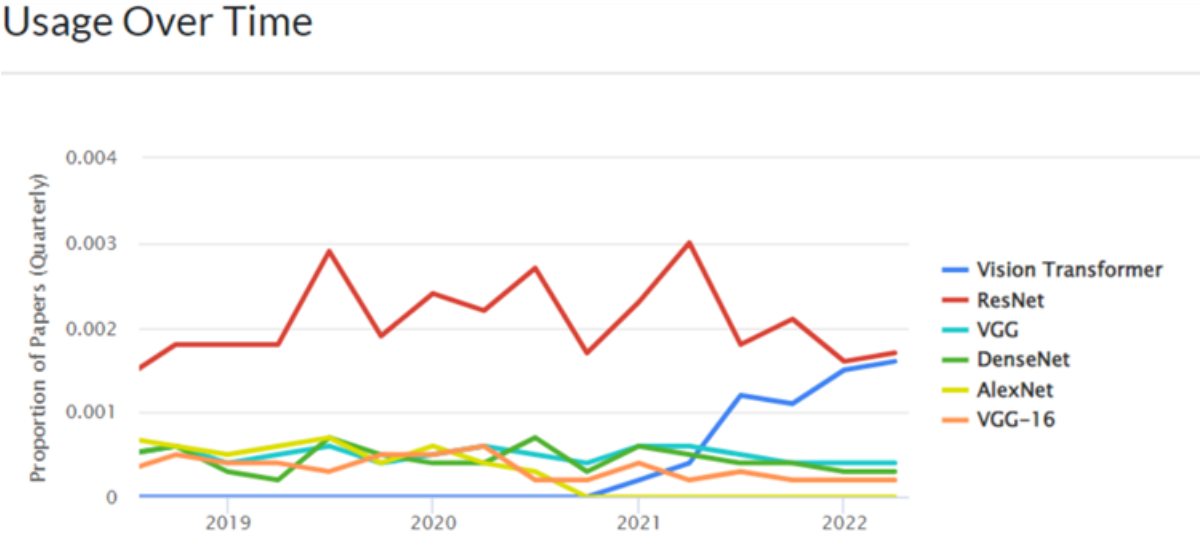


Figure 2.20: As of 2022, the use of a Vision Transformer (ViT) in picture tasks has surpassed all other common CNN architectures and matches the use of ResNets.

Brief History

1. Attention in Language Models
   A group of Google researchers published a landmark paper titled "Attention Is All You Need" in 2017 [115], which served as the catalyst for the Deep Learning Transformer revolution. It proposes a novel architecture, which models long-range dependencies in sequential data (e.g. text), by arranging a set of self-attention layers. The model utilizes a self-attention layer to concurrently focus on several input sequence components. It may be used, for instance, to determine the distance (and thus, the connection) between each word in a given phrase. The text-based Transformers BERT by Google and GPT-3 by OpenAI are examples of implementations. By 2021, BERT, among many other applications, will process and automatically fill out each and every English-based Google.com user search. The majority of uses for Transformers are found in language modeling and Natural Language Processing (NLP). As a result, they are frequently compared to RNNs, more especially the Long Short-Term Memory (LSTM) architecture [118]. In order to transmit information sequentially during the encoding and decoding of each word token, LSTMs rely on hidden states. They frequently fail to learn long-term dependence, though [119].

2. Attention in Vision Tasks
   The attention mechanism plays a pivotal role in vision tasks by enabling the network to focus selectively on specific regions or objects within the input image. Unlike Convolutional Neural Networks (CNNs), which rely

on variable-sized convolutional kernels to scan features hierarchically across multiple levels, attention mechanisms operate within a single network layer to dynamically emphasize important features [120]. As depicted in Figure 2, attention mechanisms tokenized the image at the pixel level, where each pixel interacts with every other pixel in the grid. However, this process incurs a computational cost of $O(n^2)$, where n is the width of the square image. To address this challenge, the input image is divided into equally sized square patches, referred to as image patches. These patches are then flattened into one-dimensional sequences of size n×1 and enriched with positional embeddings that encode spatial information [121]. These embeddings are input into a Transformer model, which processes them using its attention mechanism to capture both global and local dependencies. The output from the Transformer is then passed through a feed-forward classifier, typically a Multilayer Perceptron (MLP), to generate predictions in the form of a probability distribution. This architecture is especially effective for vision tasks as it overcomes the limitations of CNNs by modeling long-range dependencies and global context with greater flexibility.

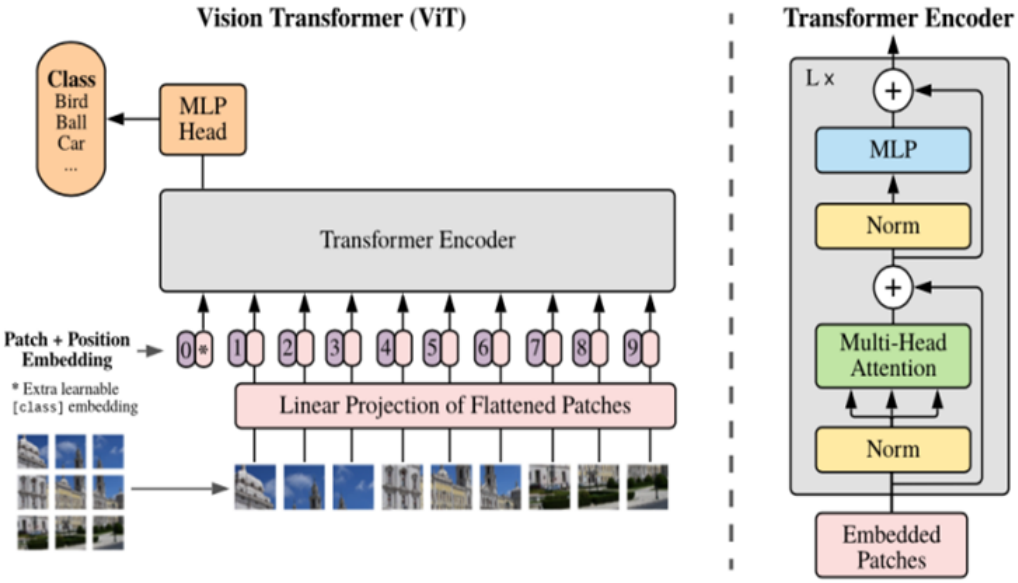


Figure 2.21: The Vision Transformer (ViT) architecture.

In a manner, a Transformer is a generalization of a feed-forward network, however unlike an MLP where connection weights are fixed, each connection weight (i.e. attention) in a Transformer. Because of this, the Transformer is permutation invariant, unlike the MLP, which means that it cannot determine the source of some information without the use of additional learnable

sequential positional embedding, such as the indexing of the picture patches [122].

Structure of Transformers

A Deep Learning architecture called The Transformer that was initially proposed in [115] is mostly based on the attention mechanism. This architecture was created to handle sequences or, more generally, data that may be separated into tokens with meaningful ordering. Transformers have therefore been used in natural language processing first before being used in other domains, such computer vision [123]. As a result of its first use in the field of machine translation, the Transformer architecture's whole structure, seen in figure 3.1, is made up of two primary components: an encoder and a decoder. Automatic translation from one language to another is carried out because the later attempts to reconstruct a new equivalent sentence from these embedding, whereas the former seeks to identify a latent representation of a source sentence. In several works [124], only the encoder element is used, which is treated as a feature extractor; the given task, typically classification, is then carried out with an MLP head that processes the extracted features. This is to be noted, though, given the great ability of this architecture to extract a powerful representation from its input. This method will also be used in this project.

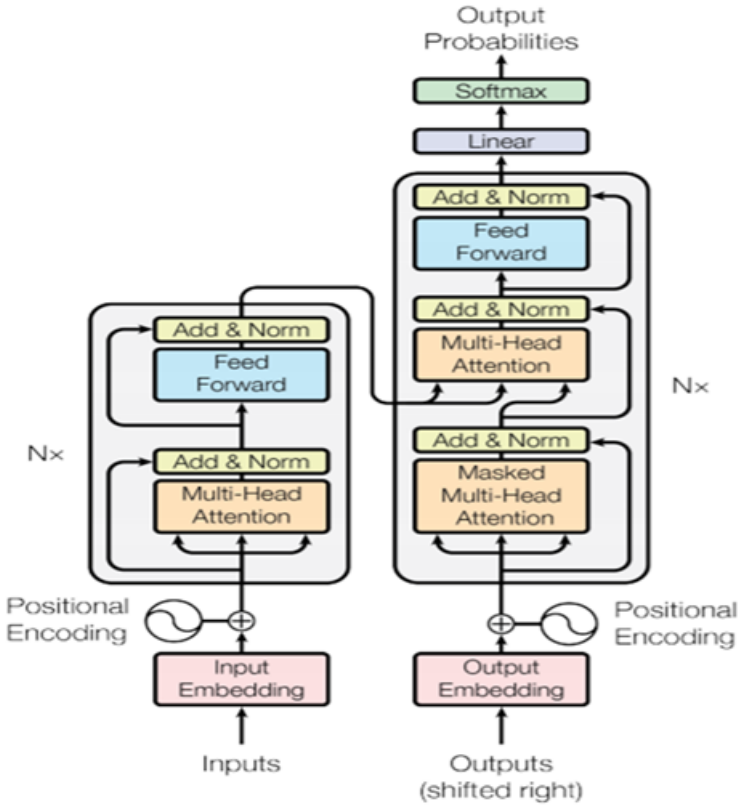


Figure 2.22: The Transformer model architecture [115]

Transformers are receiving a lot of attention due to the various benefits the attention mechanism offers:

- Higher computation parallelization as compared to RNNs. RNNs' internal stately representation, which is reached through a number of sequential operations, is one of their major bottlenecks. In a basic recurrent layer, tokens are typically processed one at a time, which means that the n-th token cannot be processed until all n-1 preceding tokens have been sequentially processed. In other words, O (n) is the number of consecutive operations needed for an input sequence containing n tokens. Instead, it is possible to calculate attention for each token separately from the others; nevertheless, this does not imply that the outcome of the attention computed for a particular token is unrelated to any other input token, as it in fact is. There is no ordering need for executing calculations, therefore their parallelization is allowed because the outcome of the attention does not directly depend on any state variable or any prior result computed from the other tokens.

- Longer range dependencies the inability to simulate long-range dependencies is a result of the constrained receptive field of CNNs and the vanishing

gradient problems of conventional RNNs. While some other advancements, including skip connections of LSTM cells in RNNs and a larger dilation value in convolutional layers, at least partially ameliorate these issues, it has been discovered that attention can represent dependencies with extraordinary efficacy in even longer sequences.

- Transfer Learning recurrent layers haven't had much success with transfer learning up to this point, thus fine-tuning an RNN that has already been trained has always been challenging and hasn't ever produced results that are noticeably better results than training from zero. It has been empirically demonstrated that attention modules and Transformers in general are simpler to fine-tune, which may be advantageous for learning a variety of tasks more successfully.

- Added interpretable models since the attention mechanism is based on computing a linear combination of all tokens, it is simpler to inspect and interpret partial results, as demonstrated in [8], where the results of attention seem to be related to the syntactic and semantic structure of the input sentences.

2.3.5 Fundamental Concepts in ViTs

The fundamental architecture of Vision Transformers (ViTs) is depicted in Figure 2.21 [20]. In this architecture, the input image undergoes preprocessing where it is split into fixed-size patches, compressed, and converted into Patch Embeddings—linear embeddings of lower-dimensional representations [125]. These embeddings are supplemented with positional embeddings and class tokens, then fed into the encoder block of the transformer to produce class labels. The encoder block comprises four main components: a feed-forward neural network (FFN), a normalization layer, a residual connection, and a multi-head self-attention (MSA) layer. The model's final output prediction is generated by the last layer, often referred to as the MLP layer or decoder block. The following subsections explain the core concepts underlying each element of the ViT architecture in detail [126].

Patch Embedding

Patch embedding is a cornerstone of the ViT architecture [21]. To adapt image data for processing by a transformer, which inherently operates on sequential data, ViTs convert input images into sequences of vector representations [20]. This is achieved by dividing the input image into fixed-size, non-overlapping patches, flattening these patches into one-dimensional vectors, and projecting them into a higher-dimensional feature space using a linear layer with D embedding dimensions.

This process is mathematically expressed as:

$$X_{patch}^{NxD} = R(I_{image}^{AxBxC}) \tag{2.1}$$

The input image is I_{image} with size $AxBxC.R(.)$ is the reshaping function to produce N number of patches X_{patch} with size D, and $N = \frac{A}{P}X\frac{B}{P}, D = PP.C$, P = patch size and C = channels This embedding method enables ViTs to learn long-range relationships between patches, which is critical for excelling in image-related tasks.

Positional Embedding

Vision Transformers utilize Positional Embeddings to encode spatial relationships between patches in an image, addressing the lack of inherent positional information in visual data [127]. Unlike natural language, where word order conveys meaning, images lack an explicit ordering of pixels, necessitating positional encoding to guide the transformer.

In this context, positional encoding is added to the input embeddings, providing the model with critical spatial information. This encoding helps the transformer understand the relative and absolute positions of patches, enabling it to capture location-dependent features effectively.

Typically, positional embeddings are represented as grids added to the input data before being passed through the transformer model. This approach allows the transformer to consider spatial arrangements and learn location-sensitive features [128].

Since the introduction of ViTs, various advancements in positional embedding techniques have been proposed to improve the model's capacity to learn spatial relationships, even though patches are processed sequentially. These advancements ensure that ViTs excel in capturing both local and global contextual features of an image.

The selection of a positional embedding technique in a Vision Transformer (ViT) architecture is influenced by multiple factors, including the specific application, the required model complexity, and the constraints on computational resources [129]. Positional embeddings play a vital role in helping the model understand the spatial relationships between patches, but the choice of method can significantly impact both the performance and efficiency of the model.

Researchers are actively investigating and proposing novel positional embedding techniques to address the limitations of existing methods, with a focus on improving the scalability and accuracy of ViTs across diverse tasks. These advancements aim

to enhance the model's ability to capture fine-grained spatial information while maintaining computational efficiency, ensuring that ViTs remain versatile and effective for a wide range of applications [130].

Attention Mechanism

The attention mechanism in Vision Transformers (ViTs) is an evolution of the self-attention mechanism originally developed for Natural Language Processing (NLP). In ViTs, the attention mechanism enables the model to focus on spatial relationships and interactions between different patches of an image. This approach differs fundamentally from Convolutional Neural Networks (CNNs), which rely on convolution kernels for feature extraction.

1. Self-attention mechanism
   The self-attention mechanism is a core component of the ViT (Vision Transformer) architecture, playing a vital role in understanding the relationships between different parts of an image[21]. Self-attention computes pairwise relationships between all tokens (image patches) to determine how much each token contributes to others. Each patch (token) interacts with every other patch, allowing the model to capture long-range dependencies and spatial relationships. This mechanism is mathematically represented as:

$$Attention(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d_k}})V. \tag{2.2}$$

   where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors.

   - Step 1: Query, Key, and Value Matrices:
     The model generates three matrices from each patch embedding:
     Query (Q): Represents what the model is "looking for" in other patches.
     Key (K): Represents the "information" other patches possess.
     Value (V): Contains the actual features of each patch.
     These matrices are created by multiplying the patch embeddings with learned weight matrices.
   - Step 2: Attention Scores:
     The model computes attention scores between each pair of patches by multiplying the query of one patch with the transpose of the key of another patch.
     These scores indicate the relevance of one patch to another based on their features.

- Step 3: Weighted Sum:
  The attention scores are then normalized using a softmax function to obtain weights.
  These weights are used to create a weighted sum of the value vectors of all patches.
  This weighted sum essentially provides a contextual representation of the current patch, incorporating information from relevant parts of the image. where the weights are calculated by a scoring function (Equation 2.3).

2. Multi-Head Self-Attention (MSA)
   Multi-Head Self-Attention (MSA) is a key component in transformer architectures, including Vision Transformers (ViTs) [121]. MSA enhances the model's ability to capture diverse and complex relationships within the input sequence or image by allowing it to attend to different parts of the input simultaneously. Here's an overview of how Multi-Head Self-Attention works [120]:

   - Single Head Self-Attention:
     In standard self-attention, the model computes attention weights for each element in the input sequence based on its relationship with every other element. The weighted sum of these elements forms the output for each position in the sequence.
   - Multiple Heads:
     In Multi-Head Self-Attention, the self-attention mechanism is applied multiple times in parallel, each with its own set of learnable parameters (weights). Each attention head produces a different set of attention weights and corresponding output.
   - Concatenation or Averaging:
     The outputs from all attention heads are typically concatenated along a specified dimension or averaged to produce the final output. This allows the model to capture different aspects of relationships and dependencies in the input sequence.
   - Linear Projection:
     After concatenation or averaging, a linear transformation is often applied to the aggregated output to project it back to the desired dimensionality. This linear projection helps the model to adaptively learn relationships in a more complex space.

   Mathematically, the operation of Multi-Head Self-Attention can be expressed

as follows [122]. Let HH be the number of attention heads, d_{model} be the model's hidden dimension, and X be the input sequence[20]. The output Y is obtained by concatenating the outputs from each attention head and applying a linear transformation:

$$Y = Concat(head1, head2, ..., headH)W_0, \tag{2.3}$$

where W_0 is the output projection matrix.

$$head_i = Attention(Q_i, K_i, V_i), where i = 1, 2, ..., h \tag{2.4}$$

Enables modeling long-range dependencies. Unlike CNNs with limited receptive fields, self-attention allows Vision Transformers (ViTs) to capture relationships between distant patches in the image [121]. Improved performance: self-attention has significantly contributed to the success of Vision Transformers (ViTs) in various computer vision tasks, achieving competitive results compared to traditional CNN architectures. Flexibility: The self-attention mechanism can be applied to various patch sizes and configurations, providing flexibility in model design [120].

Transformer layers

A Vision Transformer (ViT) encoder consists of multiple transformer layers [20]. Each transformer layer typically contains two sub-layers: a Multi-Head Self-Attention (MSA) mechanism and a Feedforward Neural Network (FNN) layer [120]. The encoder stack's layers are organized as follows:

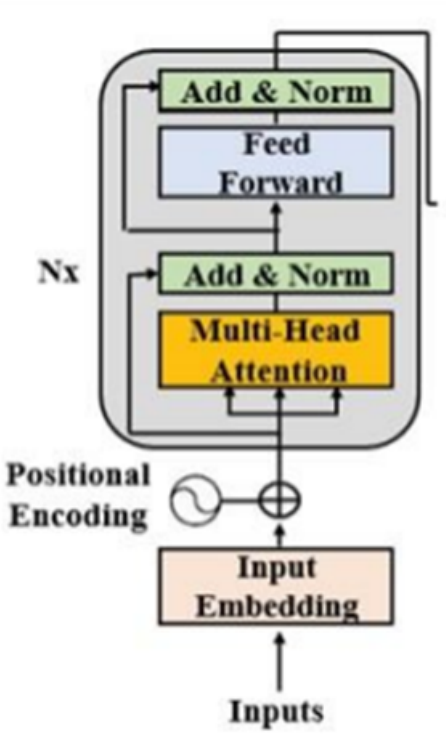


Figure 2.23: The model encoder structure of a transformer layer

The entire N=6 layers of the Transformer model retain the original encoder layer structure. A multi-headed attention mechanism and a fully connected position-wise feed forward network are the two main sub-layers that contain each layer [131]. Here's a high-level overview of the transformer layers in a ViT encoder:

- Feed-Forward Network (FFN):
  Following the MSA, a feed-forward network is typically applied independently to each position in the sequence.
  The FFN consists of two linear transformations separated by a non-linear activation function (commonly a ReLU activation).
  The output of the FFN is calculated as follows:

$$FFN(X) = ReLU(X.W_1 + b_1).W_2 + b_2 \tag{2.5}$$

  where X is the input, $W_1, b_1, W_2, and b_2$ are learnable parameters.

- Layer Normalization and Residual Connection:
  Both the MSA and FFN are typically followed by layer normalization, which helps stabilize and normalize the activations [132].
  The output of each layer is then obtained by adding a residual connection from the input to the normalized output. Mathematically, this is represented as

$$LayerOutput = LayerNorm(X + Sublayer(X)) \tag{2.6}$$

where $Sublayer(X)$ represents either the MSA or FFN operation.

- Position-wise Feed-Forward Networks: In transformers, including ViTs, the term "position-wise" refers to the fact that the FFN is applied independently to each position in the sequence. This is in contrast to "channel-wise" operations, which would involve operations across the entire sequence but within a specific channel [133].

Each layer within a Vision Transformer (ViT) encoder is composed of four fundamental components [134]. Multi-Head Self-Attention (MHA): This mechanism enables the model to compute pairwise relationships between patches (tokens) in an input image, capturing dependencies and interactions across different regions. The multi-head design allows the model to attend to various aspects of the input simultaneously, enhancing its ability to learn diverse patterns and features. Feed-Forward Network (FFN): Positioned after the attention mechanism, the FFN applies non-linear transformations to the output of the MHA. This two-layer network (often with a GELU activation function) processes the information at each token independently, enriching the representation with higher-level abstractions. Layer Normalization: Normalization ensures stable training by keeping the outputs of each layer well-scaled and reducing the risk of exploding or vanishing gradients. It is applied both before and after the MHA and FFN components in some implementations. Residual Connections: To combat the degradation of performance in deeper networks, residual (or skip) connections directly add the input of a layer to its output. This facilitates the flow of gradients during backpropagation, aiding in convergence and improving model stability.

2.4 Conclusion

Facial beauty prediction, leveraging deep learning techniques, offers flexibility in algorithm selection depending on the desired application. One can opt for either supervised or semi-supervised methods. While both paradigms are explored in the literature, supervised methods—particularly those relying on Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs)—dominate in the domain of facial beauty prediction. CNNs are uniquely adept due to their mathematical architecture, which is tailored to efficiently address image-related challenges. They excel in tasks requiring classification and alignment by capturing spatial hierarchies in visual data. Furthermore, CNNs remain dominant in scenarios where datasets are limited, and high-resolution image inputs are not mandatory. They have become the most popular solution for computer vision tasks in facial beauty prediction

due to their: Proven reliability and accuracy. Extensive research history and optimization. Ability to handle various data representations effectively.

ViTs represent a promising new direction for computer vision. Unlike CNNs, ViTs use a transformer-based architecture that excels in modeling long-range dependencies and capturing global relationships within an image. Some of the key advantages of ViTs include: Their ability to process entire images as sequences, enabling holistic feature extraction. Superior performance on large datasets where their scalability shines.

However, ViTs are still a relatively nascent technology, and the body of research surrounding their application to facial beauty prediction is not as extensive as that of CNNs. Despite these challenges, ViTs are likely to gain traction as they continue to evolve and improve.

While CNNs currently dominate the field of facial beauty prediction due to their robustness and versatility, ViTs hold the potential to revolutionize this space as they become more refined and widely adopted. Their distinct strengths complement the capabilities of CNNs, making both architectures valuable tools depending on the data and task requirements. As research into ViTs progresses, they are poised to play a crucial role in advancing facial beauty prediction methodologies.

Chapter 3

Proposed Approaches

3.1 Introduction

Facial beauty prediction is a task within the broader field of computer vision and artificial intelligence that aims to develop algorithms capable of assessing or predicting the perceived beauty of human faces.
The science behind facial beauty metrics involves the analysis of various facial features and their relationship to perceived attractiveness. Our resarch have found that certain facial features, such as symmetry, averageness, and facial proportions, play a significant role in determining facial attractiveness. By utilizing computer vision algorithms and machine learning techniques, researchers can extract these facial features from images and quantify their impact on beauty ratings. By training deep learning models on datasets, we can create predictive models that accurately estimate facial beauty based on these identified features. With the use of deep learning techniques such as convolutional neural networks and vision transfomers, Our contributions have been able to analyze facial features and accurately predict facial beauty ratings.

In this chapter, we presented various proposed approaches of taxonomy studied in the thesis as shown in Fig. 3.1. This chapter is concerned with describe the proposed methods for facial beauty prediction. The first algorithm we proposed uses CNN (see Section 2).

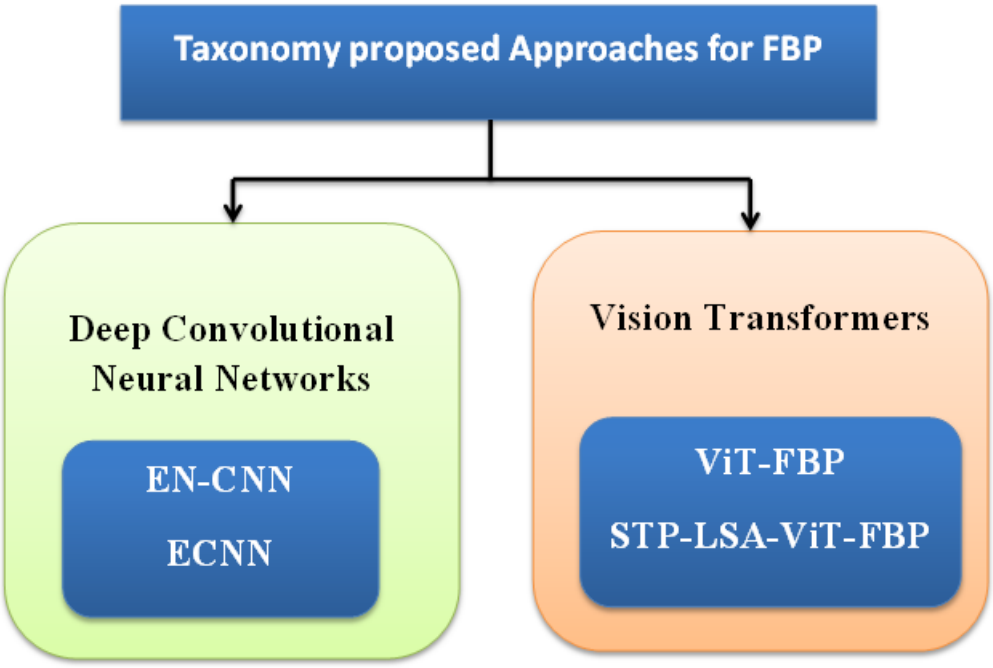


Figure 3.1: Categories of the proposed approaches studied in the thesis.

3.2 Proposed approaches using DCNN

3.2.1 First proposed approach

This study introduces an ensemble of trained models (EN-CNN) as an advanced framework for facial beauty prediction. The proposed EN-CNN model leverages transfer learning (TL) with four pre-trained architectures: DenseNet201, InceptionV3, MobileNetV2, and EfficientNetB7. These models are fine-tuned for a regression task and combined to generate superior predictions of facial beauty. The ensemble approach capitalizes on the strengths of each individual model to achieve high accuracy and robustness.

1. Pre-trained deep learning models with transfer learning (TL)
   Transfer learning is a technique that adapts models trained on one task for another related task. Pre-trained models, developed using extensive datasets like ImageNet, are fine-tuned or used as feature extractors for specific tasks. TL reduces training time and resource requirements while enhancing performance by transferring knowledge from the base task to the target task [135] [136].

2. DenseNet201
   DenseNet201 is a pre-trained deep learning model from the Dense Convolutional Network (DenseNet) family, proposed by Huang et al. in 2017 [137]. It

builds on the core idea of enhancing feature reuse and reducing redundancy by densely connecting each layer to every other layer in a feed-forward fashion. DenseNet201, as its name implies, is a version of DenseNet with 201 layers. DenseNet is built on ResNet's design but introduces dense connections that share extracted features across layers. This approach enhances feature propagation and reduces redundancy. DenseNet is composed of composite layers with Rectified Linear Unit (ReLU) activation and fully connected layers, making it highly effective when fine-tuned via transfer learning for facial beauty prediction.

3. Pre-trained InceptionV3
   InceptionV3 [138], introduced by Google's team, builds upon the InceptionV1 model with enhancements like smaller convolution filters (e.g., 1×7 and 1×5) and 1×1 convolutions for feature compression[139]. By combining different filter sizes in parallel, InceptionV3 achieves improved feature representation and computational efficiency, making it a powerful tool for vision tasks[140].

4. Pre-trained MobileNetV2
   MobileNetV2 [141] utilizes inverted residuals and linear bottlenecks, making it a lightweight CNN optimized for low-latency and low-power applications[142]. With components like depthwise separable convolutions (DSC), MobileNetV2 significantly reduces computational complexity, using only 22% of the resources required by traditional CNNs [143]. Its design is ideal for hardware-constrained environments, maintaining efficiency without sacrificing accuracy.

5. Pre-trained EfficientNetB7
   Pre-trained EfficientNetB7 [144] refers to a pre-trained deep learning model based on the EfficientNet architecture[145], specifically the B7 variant. EfficientNet [145]is a family of convolutional neural networks (CNNs) designed to achieve high accuracy while maintaining computational efficiency. It achieves this by using a compound scaling method that balances network depth, width, and resolution. EfficientNet models come in different variants, denoted by letters A to B7, with B7 being one of the larger and more powerful models in the family .By leveraging the pre-trained weights from ImageNet, EfficientNetB7 can learn faster on new tasks compared to starting from scratch. The pre-trained features can act as a strong foundation for learning task-specific features, potentially leading to better performance on various computer vision tasks like image classification, object detection, or image segmentation [146] . Utilizing a pre-trained model can significantly reduce the computational resources required for training, especially for tasks with

limited datasets. Pre-trained EfficientNetB7 is a powerful tool for various computer vision tasks. Its pre-trained weights and efficient architecture offer advantages in terms of training speed, performance, and resource usage.

6. The proposed EN-CNNs
   The EN-CNNs architecture combines the predictions of the four pre-trained models to improve accuracy and reduce biases inherent in individual architectures. An 80-20 split was used, with 80% of the data for training and 20% for testing. Optimizer: The Adam optimizer was employed for its efficient calculations, minimal memory requirements, and ability to handle noisy gradients. The ensemble combines the strengths of DenseNet201, InceptionV3, MobileNetV2, and EfficientNetB7, leveraging their diverse architectures and feature extraction capabilities. The ensemble ability to aggregate predictions from multiple models provides a robust framework for automated facial beauty prediction, delivering superior performance in terms of efficiency and accuracy. Figure 3.2 showcases the EN-CNN workflow, illustrating the data flow and integration of pre-trained models in the ensemble. This approach underscores the value of leveraging state-of-the-art architectures for complex vision tasks like facial beauty prediction.

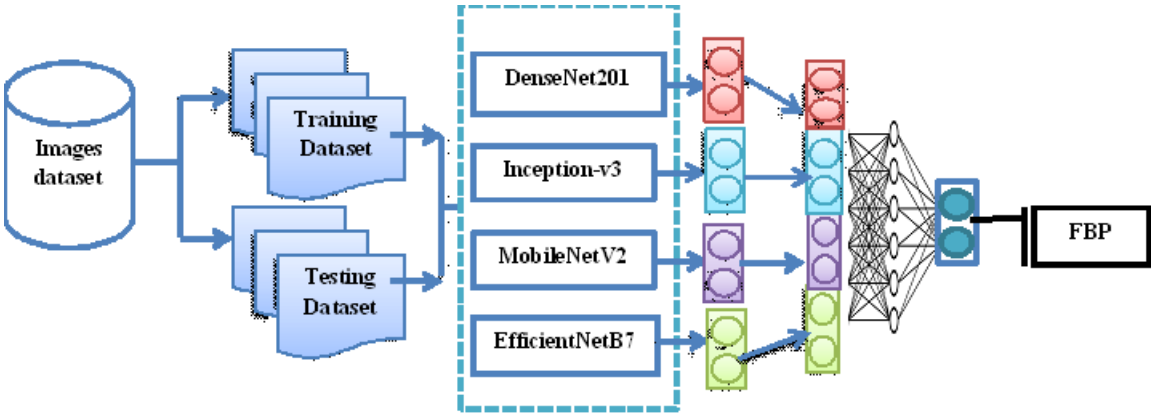


Figure 3.2: The proposed deep CNN ensemble networks (EN-CNNs)

The EN-CNN approach demonstrates the potential of combining diverse deep learning architectures in predicting facial beauty. With optimized transfer learning and ensemble techniques, the model achieves high performance while addressing challenges like overfitting and limited data availability. Below is algorithm 1 outlining the methodology:

Algorithm 1 Algorithm 1 represents the instructions used to compute the ensemble final prediction during image testing.

Input models: M=[DenseNet201, InceptionV3, MobileNetV2 and EfficientNetB7 models]

Data: SCUT-FBP5500 dataset

1 **for** *i in A* **do**

2 **for** *j in M* **do**

3 k[I,j]= ▷ prediction for the test image i when using j ▷ Compute the ensemble final prediction for test image i based on mean and k[j,:] (EN-CNN(Mean))

Output: prediction for A using (EN-CNN(Mean))

3.2.2 Second proposed approach

This study introduces an innovative approach to facial beauty prediction using advanced deep learning techniques. The central focus is on developing a robust ensemble regression model that combines the strengths of multiple architectures to accurately estimate facial beauty scores. Investigating the effectiveness of transfer learning for facial beauty prediction. Pre-trained models like InceptionV3 and MobileNetV2 are fine-tuned to adapt their learned features for the specific task of predicting facial beauty. Integrating InceptionV3, MobileNetV2, and a newly proposed S-CNN (Simple CNN) architecture into a three-branch ensemble network. Training each network using tailored loss functions to optimize their individual contributions to the final prediction. Designing an ensemble method that aggregates the predictions of the three models to produce a final facial beauty score. This ensemble approach aims to leverage the complementary strengths of the individual networks, improving accuracy and robustness. Fine-tuning hyperparameters (e.g., learning rates, batch sizes, and model-specific settings) to maximize the performance of pre-trained models for the classification and regression tasks involved in beauty prediction. Enhancing the alignment of machine-predicted beauty scores with human judgment, ensuring that the predictions reflect subjective aesthetic evaluations. Outperforming existing baseline methods by achieving higher prediction accuracy and better correlation with human evaluations. The E-CNN model demonstrates its potential as a state-of-the-art solution in facial beauty prediction tasks. This approach not only pushes the boundaries of facial beauty prediction accuracy but also highlights the practicality of leveraging deep learning ensembles for nuanced human-centered tasks.

1. S-CNNs network

   The S-CNNs (Simple Convolutional Neural Networks) architecture is a

lightweight, efficient model designed to predict facial beauty. The key features and contributions of this architecture are detailed below:

- Architecture Design Convolutional Layers: Multiple 2D convolution layers form the backbone, each performing spatial feature extraction. The layers are followed by the Rectified Linear Unit (ReLU) activation function, which introduces non-linearity into the model and accelerates convergence. Fully Connected Layer: A final fully connected layer integrates the extracted features into a unified representation to produce the prediction score.
- MixConv Integration MixConv (Mixed Convolution): The S-CNNs architecture employs MixConv techniques to combine convolutional filters of varying kernel sizes (e.g., 3×3, 5×5, and 7×7) within the same layer. Layer Mixing: By leveraging kernels of different sizes, the model can capture features at multiple scales, enhancing its ability to process diverse facial characteristics.
- Innovative Contributions

  Multi-Scale Feature Extraction: Using mixed kernel sizes (3×3, 5×5, and 7×7), the model captures both fine-grained details (e.g., textures) and global patterns (e.g., facial proportions).

  Ensemble of Simple CNN Models: Instead of relying on a single architecture, S-CNNs integrate multiple simple CNN models through ensemble techniques. This boosts robustness and reduces overfitting on the facial beauty prediction task.

  Efficiency and Lightweight Design: The architecture is computationally efficient and suitable for deployment in scenarios with limited resources while maintaining competitive performance.
- Key Advantages Flexibility: The use of MixConv allows S-CNNs to adapt to varying feature complexities within facial images. Compactness: The minimal design ensures a balance between computational overhead and predictive accuracy. Improved Accuracy: The architecture's ability to combine features across scales leads to more reliable beauty prediction.

The architecture, as indicated in Figure 3.3, highlights: Multiple convolutional layers with ReLU activation. Diverse kernel sizes for feature extraction. A final fully connected layer for regression.

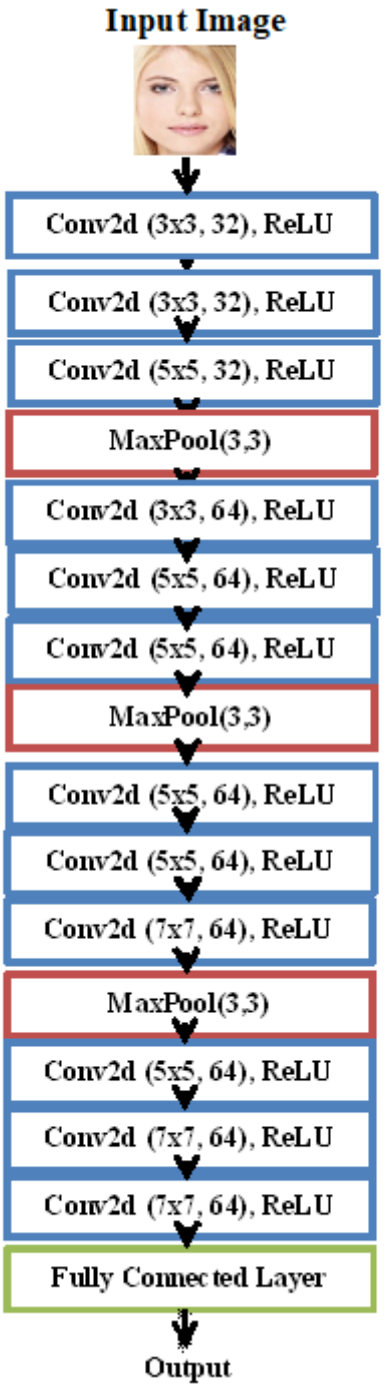


Figure 3.3: The architecture of S-CNNs network

The SCNNs network forms an essential component of the proposed ensemble model (ECNN), contributing unique multi-scale features and enhancing the overall accuracy of facial beauty predictions. Its innovative design ensures that it can serve as a standalone predictor while complementing other sophisticated networks like InceptionV3 and MobileNetV2 in the ensemble.

2. InceptionV3
   InceptionV3 is part of the Inception family of models, known for their

modular architecture called "Inception modules." These modules aim to improve computational efficiency and capture multi-scale information. Inception Modules combines convolutions with different filter sizes (1x1, 3x3, 5x5) and pooling operations to capture features at multiple scales. Auxiliary classifiers is intermediate classifiers that help in combating the vanishing gradient problem and regularize the network. figure 3.4 illustrate the inception v3 module.

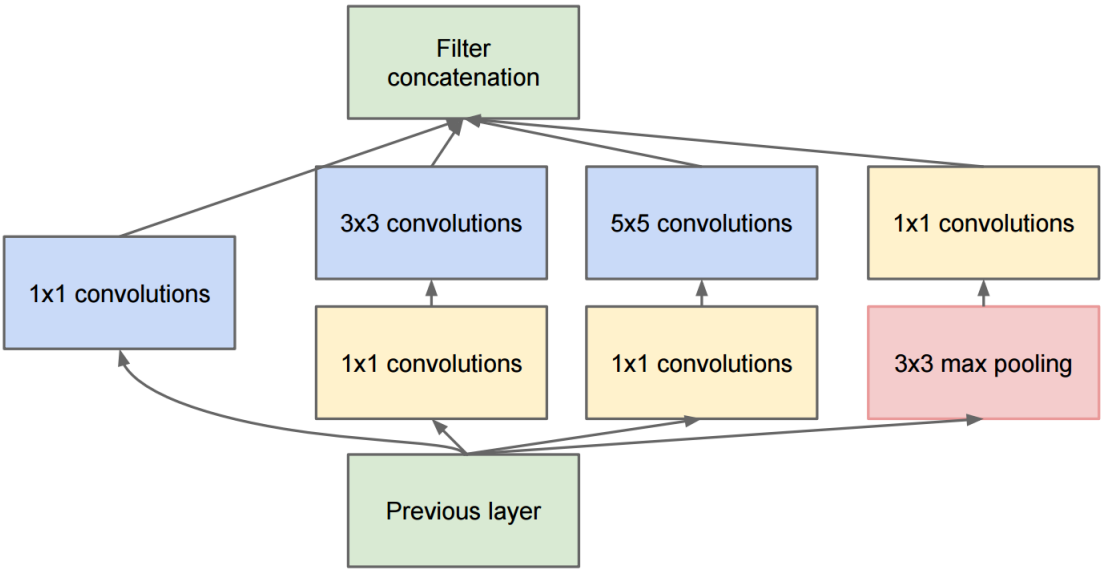


Figure 3.4: The architecture of InceptionV3 module

$$y = [conv1x1(x), conv3x3(x), conv5x5(x), pool(x)] \quad (3.1)$$

Concatenating the outputs of different convolution and pooling operations.

3. MobileNetV2
   MobileNetV2 is designed for mobile and edge devices, focusing on lightweight and efficient models. It introduces a novel layer called the "inverted residual with linear bottleneck." Depthwise separable convolutions is reduces the number of parameters and computation by splitting the convolution into depthwise and pointwise operations. Inverted residuals use the input and output are thin bottleneck layers, while the inner layer is an expanded high-dimensional space.figure 3.5 illustrate the inception v3 module

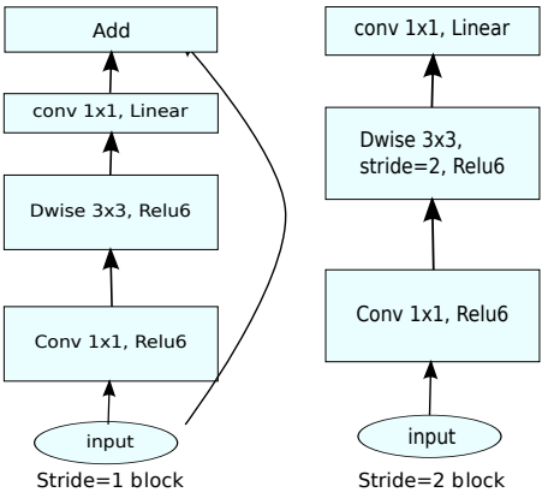


Figure 3.5: The architecture of MobileNetV2 module

4. The proposed E-CNNs The proposed ensemble of deep CNNs (E-CNNs) architecture follows three previously trained models: InceptionV3, MobileNetV2, and S-CNN. In the proposed E-CNNs for the automated classification system, they serve as fundamental classifiers of facial beauty. The details of the proposed E-CNNs are as follows:

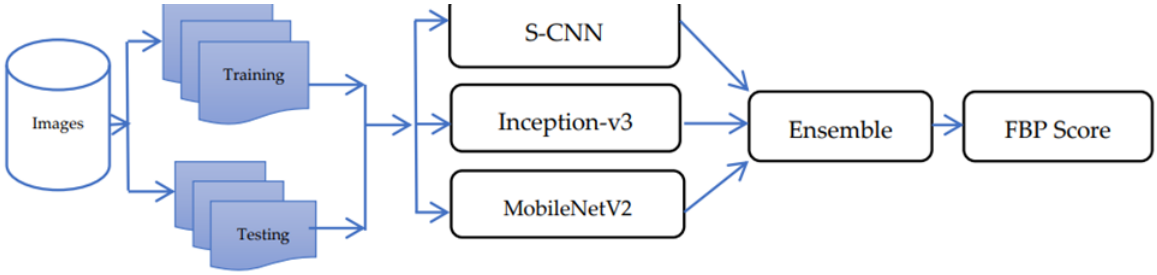


Figure 3.6: The proposed deep CNN ensemble networks (E-CNNs)

The E-CNNs (Ensemble Convolutional Neural Networks) model is a deep learning ensemble system combining three key architectures: InceptionV3, MobileNetV2, and a newly proposed S-CNN (Simple CNN). These networks are integrated to enhance the performance and accuracy of facial beauty prediction through an ensemble strategy. Below is a detailed breakdown of the proposed E-CNNs algorithm and its methodology: Algorithm Steps

- Step 1: Preprocessing

Data Preparation: Load and preprocess facial images. Normalize pixel values to meet model input requirements (e.g., range [0, 1]). Resize

images to the required dimensions for the three architectures (e.g., 224×224×3). Split data into training (80%) and testing (20%) subsets.

Data Augmentation: Apply techniques like flipping, rotation, scaling, and cropping to improve model robustness.

- Step 2: Individual Model Training

  InceptionV3 Model: Use the modular structure of InceptionV3, including its multi-scale filter sizes (1×1, 3×3, 5×5). Train using transfer learning, freezing the base layers and fine-tuning the top layers for facial beauty prediction.

  MobileNetV2 Model: Leverage its lightweight architecture with depth-wise separable convolutions and inverted residuals. Train and fine-tune using a similar transfer learning strategy.

  S-CNN Model: Construct a simple CNN using a mix of convolutional layers with kernel sizes 3×3, 5×5, and 7×7, followed by ReLU activations. Add a fully connected layer for regression at the end. Optimize the S-CNN architecture with MixConv for better feature learning.

- Step 3: Ensemble Model Construction

  Prediction Functions: Let f_1, f_2, f_3 represent the prediction functions for InceptionV3, MobileNetV2, and S-CNN, respectively.

  Individual Predictions: For an input image x, the models generate predictions: $y_1 = f_1(x)$, $y_2 = f_2(x)$, $y_3 = f_3(x)$ Ensemble Prediction: Combine the individual predictions using a soft voting strategy (arithmetic mean):

$$y_{ensemble} = \frac{1}{N}\Sigma_{i=1}^{N} y_i \tag{3.2}$$

  Here, N=3 (number of models).

- Step 4: Optimization

  Loss Function:
  Use Mean Squared Error (MSE) as the loss function to minimize the error between predicted $y_{ensemble}$ and actual beauty scores y_{true}.
  Optimizer:
  Use the Adam optimizer for efficient weight updates, benefiting from adaptive learning rates and momentum. Evaluation Metrics:
  Evaluate performance using: Pearson Correlation Coefficient (PCC):

To measure alignment with human judgment. MAE (Mean Absolute Error): To assess prediction accuracy.

This innovative ensemble framework represents a robust approach to facial beauty prediction, optimizing deep learning architectures for higher precision and broader applicability.

The research on facial beauty prediction using deep convolutional neural networks has several implications in the field of facial recognition technology and beauty analysis:

- Enhanced Facial Beauty Prediction: The development of ensemble regression models using deep CNNs can lead to more accurate predictions of facial beauty. This can have applications in various industries such as cosmetics, plastic surgery, and entertainment where facial attractiveness plays a significant role.

- Improved Human-Computer Interaction: By aligning beauty assessment with human judgment, the research can contribute to enhancing human-computer interaction experiences. Systems that can accurately predict facial beauty may be integrated into applications for virtual try-on experiences, personalized recommendations, and more.

- Advancements in Transfer Learning: The study highlights the effectiveness of transfer learning techniques in the context of facial beauty prediction. This can pave the way for further research and applications of transfer learning in other areas of computer vision and deep learning.

- Potential for Personalized Beauty Solutions: The accurate prediction of facial beauty using deep CNNs opens up possibilities for personalized beauty solutions tailored to individual preferences. This can revolutionize the beauty industry by offering customized recommendations and treatments based on facial features.

- Contribution to Facial Recognition Research: The research contributes to the growing body of knowledge in facial recognition technology by exploring the application of deep learning models in predicting facial attractiveness. This can lead to advancements in facial recognition systems and algorithms.

Overall, the implications of this research extend to various domains where facial analysis and beauty assessment are relevant, offering new insights and possibilities for leveraging deep learning techniques in understanding and predicting facial beauty. In the next section, we will discuss the new techniques Vision Transformers (ViTs), which represent a significant advancement in computer vision due to their

ability to achieve high accuracy and capture long-range dependencies. While computational cost remains a challenge, ongoing research is addressing this aspect, making Vision Transformers (ViTs) a promising architecture for future facial beauty prediction applications.

3.3 Proposed Algorithms using Vision Transformer

Vision Transformers (ViTs) are a groundbreaking adaptation of the Transformer architecture, originally designed for Natural Language Processing (NLP), now tailored for computer vision tasks. Their innovative design has propelled them to achieve state-of-the-art results across a variety of computer vision domains, including image classification, object detection, and segmentation.

- Key Differences Between Transformers and Vision Transformers

  Input Representation:
  In NLP: Transformers process text sequences, typically represented as a sequence of words or subword tokens. In ViTs: Instead of words, images are split into fixed-size patches (e.g., 16×16 pixels). Each patch is flattened into a vector, and these vectors form a sequence analogous to text tokens in NLP.

  Attention Mechanism:
  In NLP: Attention models relationships between words in a sequence, capturing contextual dependencies. In ViTs: Attention models relationships between image patches, enabling the network to learn global context and long-range dependencies within the image.

1. Advantages of ViTs

   ViTs have a number of advantages over other types of computer vision models, such as convolutional neural networks (CNNs).

   - ViTs can learn long-range dependencies more effectively than CNNs. This is because ViTs use attention, which allows them to model the relationships between patches in an image, regardless of their distance apart.
   - ViTs are more efficient to train than CNNs. This is because ViTs can be trained on smaller images than CNNs.
   - ViTs are more generalizable than CNNs. This is because ViTs do not make any assumptions about the structure of the data.

2. Disadvantages of ViTs ViTs also have some disadvantages.

- ViTs can be more computationally expensive to train and deploy than CNNs. This is because ViTs use attention, which is a more complex operation than convolution.
- ViTs can be less accurate than CNNs on some computer vision tasks. This is because ViTs are still a relatively new technology, and there is less research has been done on ViTs than on CNNs.

3. The encoder of a Vision Transformer

   The encoder of a Vision Transformer (ViT) is a stack of self-attention layers that learn the relationships between the patches in an image. Before being fed to the encoder, the image is first divided into a sequence of patches. Each patch is then flattened and projected into a lower-dimensional vector. These vectors are then passed to the first self-attention layer of the encoder. The self-attention layer computes the attention weights for each patch in the sequence by taking the dot product of the patch vector with the patch vectors of all the other patches in the sequence. The attention weights are then scaled and normalized using a softmax function. The weighted sum of the patch vectors is then computed using the attention weights. This weighted sum is the output of the self-attention layer. The output of the first self-attention layer is then passed to the next self-attention layer, and so on. This process is repeated until the encoder has generated a final hidden representation for each patch in the image. The final hidden representations are then passed to the decoder of the ViT, or they can be used directly for classification or other tasks. The encoder will repeat this process for each patch in the image, generating a final hidden representation for each patch. These final hidden representations are then passed to the decoder of the ViT, or they can be used directly for classification or other tasks. The encoder of a ViT is a powerful tool that allows ViTs to learn the relationships between the patches in an image. This makes ViTs well-suited for a variety of computer vision tasks, such as image classification, object detection, and image segmentation.

3.3.1 Third proposed approach

Our proposed ViT-FBP (Vision Transformer for Facial Beauty Prediction) architecture builds on the standard Vision Transformer (ViT) framework by incorporating key enhancements to better suit the task of facial beauty prediction [148]. This section details the structure, modifications, and processing pipeline of the model. The ViT-FBP architecture utilizes 8 transformer layers for fundamental feature extraction. Two additional fully connected (FC) layers, which enhance the network's

capacity for more nuanced feature representation, leading to improved performance. Images undergo transformations such as flipping, cropping, rotation, and color adjustments to increase dataset diversity and prevent overfitting. The input image is divided into fixed-size patches (e.g., P×P), which are flattened and encoded into vector. Each vector represents the feature embedding of the corresponding patch.The core network retains the foundational structure of the original Vision Transformer, consisting of: Data Normalization Layer:
Standardizes the input data to ensure stable training.

Multi-Head Self-Attention Layer:
Captures long-range dependencies between image patches to model spatial relationships effectively.

Skip Connections:
Introduced after encoding patches to enhance gradient flow and maintain spatial consistency.

Feedforward Block:
Includes a multi-layer perceptron (MLP), which refines the feature representations.

GELU Activation Function:
Used in the MLP for smoother and efficient gradient updates.

Layer Normalization:
Further stabilizes the training process by normalizing intermediate feature representations. Regression Token a specific token in the sequence, appended during patch tokenization, is used as a global representation of the image. This token is further refined through the transformer layers. Multilayer Perceptron (MLP) Head processes the regression token to extract essential features. Two Fully Connected Layers additional FC layers further refine the extracted features, leveraging the higher capacity of the expanded network for accurate beauty score prediction. Figure 3.7 illustrates the flow of the ViT-FBP architecture, highlighting: The image tokenization layer dividing images into patche. Transformer blocks comprising multi-head attention, skip connections, and MLP. The MLP head with two fully connected layers processing the regression token for final prediction.

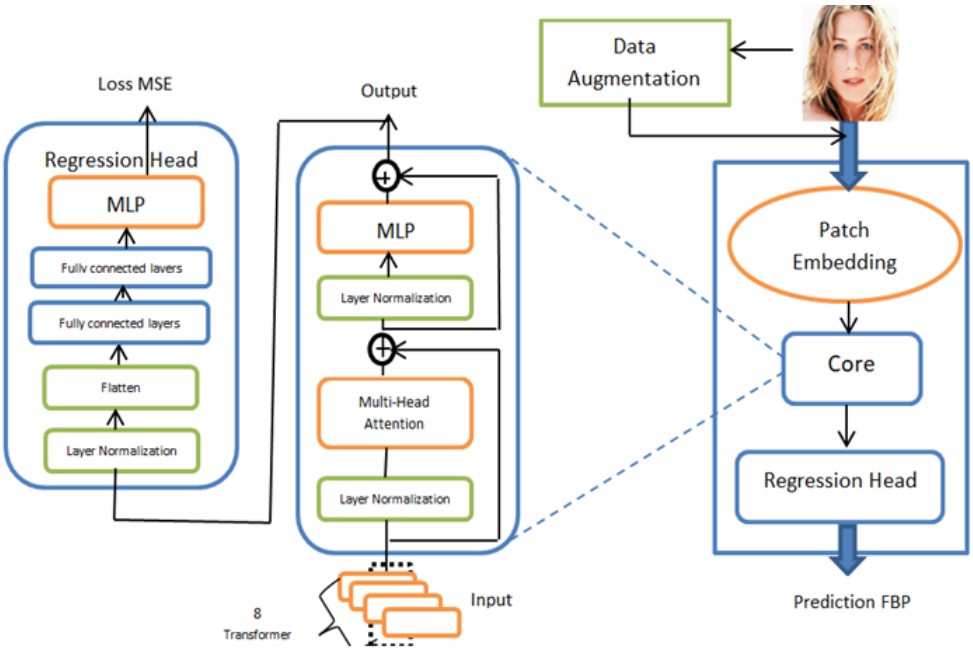


Figure 3.7: Algorithm architecture, the core block consists transformer of an MHA, an MLP, skip connections, and layer normalizations.

The ViT-FBP methodology follows a structured pipeline for leveraging Vision Transformer architecture in predicting facial beauty scores. Below is a detailed breakdown of the methodology:

- Data Preparation and Augmentation

  Dataset Preprocessing:
  Collect and preprocess face image data to ensure consistent quality and resolution. Resize all images to a standard size compatible with the model (e.g., 224×224).

  Data Augmentation:
  Apply various augmentation techniques to increase dataset variability and prevent overfitting: Geometric transformations: Flipping, rotation, and cropping. Color adjustments: Brightness, contrast, and saturation. Random noise addition: To simulate real-world variations.

- Image Tokenization

  Patch Division: Split the input image into non-overlapping patches of fixed size P×PP×P. Example: For an image of size 224×224224×224, and patch size 16×1616×16, this results in 14×1414×14 patches.

  Patch Embedding: Flatten each patch into a 1D vector and pass it through a linear projection layer to form patch embeddings.

Position Encoding: Add positional information to each patch embedding to preserve spatial relationships.

- ViT Core Architecture

  Input to Transformer Layers:
  Feed the sequence of patch embeddings (including the regression token) into the Vision Transformer.

  Transformer Block Operations:
  Multi-Head Attention: Computes attention weights between all patches, capturing long-range dependencies. Layer Normalization: Stabilizes training by normalizing input features. Feedforward Neural Network:
  Processes the attention output using a Multilayer Perceptron (MLP) with two linear layers and GELU activation. Skip Connections:
  Ensure smooth gradient flow and facilitate learning in deep networks.

  Regression Token:
  A dedicated token is used as a global representation of the image. This token aggregates information from all patches.

- Feature Extraction and Regression

  MLP Head:
  After processing by the transformer blocks, the regression token is fed to an MLP head.

  Additional Fully Connected (FC) Layers:
  Two FC layers refine the extracted features: FC Layer 1: Reduces dimensionality and extracts salient patterns. FC Layer 2: Outputs the final facial beauty score.

- Model Training

  Loss Function:
  Use a regression loss function such as Mean Squared Error (MSE) to minimize the difference between predicted and actual beauty scores. Optimization:

  Train the network using optimizers like Adam or SGD with an appropriate learning rate schedule. Regularization:

  Apply dropout and weight decay to reduce overfitting.

The multi-head attention consists of several self-attention blocks to capture the complex interactions among the different items in the sequence. In essence, we repeatedly cycle through the attention process [149]. With multi-dimensional keys, values, and queries, it is advisable to use multiple attention functions instead of

just one. This allows for executing the attention function simultaneously on each of the projected versions of queries, keys, and values. Each matrix is multiplied by a distinct weight matrix to create the mapping [150].
The dense layer inside fully connected layers consists of 2048 nodes, while the second dense layer inside fully connected layer blocks only consists of 1024 nodes . For regression models, PC improvement depends not only on the model's structural design but also on the loss functions used. During training batches, the loss function calculates the total error and utilizes backpropagation to adjust the weights. To address various domains, several loss functions have been developed, some of which are derived from existing loss functions. The loss functions also take into consideration the imbalances in the dataset.
In the case of the regression model of FBP, the default and most frequently used option is Mean Squared Error (MSE).

The ViT-FBP model utilizes a transformer architecture that processes images as sequences of patches. The input image is divided into N patches, each represented as a vector. The patches are then linearly embedded into a sequence of tokens.

$$X = Flatten(P_i).W_e \tag{3.3}$$

Where:
P_i is the i-th image patch.
W_e is the embedding matrix.
X is the resulting sequence of embedded patches.

Multi-Head Self-Attention the core of the transformer architecture is the multi-head self-attention mechanism, which allows the model to focus on different parts of the input sequence simultaneously. The attention scores are computed using the following equations:

$$Attention(Q, K, V) = softmax(\frac{QK_T}{\sqrt{d_k}})V \tag{3.4}$$

Where:
Q (queries), K (keys), and V (values) are derived from the input embeddings.
$d_k k$ is the dimension of the keys, used for scaling.
In the context of facial beauty prediction, a regression token is used to aggregate information from the attention mechanism to predict beauty scores. This token is processed through a Multi-Layer Perceptron (MLP). Output=MLP(Regression Token). Where the MLP consists of two fully connected layers with activation functions (e.g., GELU). The proposed vision transformers (ViT-FBP) are proposed

as a more effective alternative to traditional convolutional neural networks (CNNs) like ResNet for facial beauty prediction. Performance of the experimental results indicate that the ViT-FBP model consistently outperforms previously published CNN approaches, including ENCNN, ResNet, AlexNet, and others, on various facial beauty prediction tasks. Specifically, the ViT-FBP achieved a Pearson Coefficient of 0.9534 on the SCUT-FBP5500 dataset, demonstrating its superior predictive capabilities compared to traditional CNNs. Feature extraction while traditional CNNs rely on feature extraction through convolutional layers, the vision transformer approach leverages an attention mechanism that allows for better modeling of long-range dependencies in the data. This capability enhances the model's understanding of complex facial features and their contributions to perceived beauty. Data efficiency the research addresses the challenge of overfitting and the need for large datasets, which are common issues with CNNs. The vision transformer framework is proposed to tackle these difficulties, particularly when working with smaller datasets, making it a promising option for facial beauty prediction tasks. Interpretability of study suggests that the vision transformer model is not only more accurate but also more interpretable than previous methods, which is crucial for applications in fields like cosmetic surgery and aesthetic assessments.

The proposed ViT-FBP model leverages the strengths of vision transformers, including attention mechanisms and efficient feature extraction, to improve facial beauty prediction. The use of regression tokens and careful evaluation through metrics like MSE and PC further enhances the model effectiveness in this domain. The research demonstrates that vision transformers can outperform traditional CNNs in this challenging task, providing a more accurate and interpretable approach to assessing facial beauty.

The enhancements made to the ViT-FBP architecture, including additional fully connected layers and refined patch encoding, aim to better capture the nuances of facial features and improve the accuracy of facial beauty prediction. By integrating these advancements, the model achieves superior performance while maintaining the core strengths of the Vision Transformer framework.

3.3.2 Fourth proposed approach

The Transformer architecture's application to image classification, as embodied by Vision Transformers (ViTs), represents a significant advancement in deep learning for visual tasks. Below is an explanation of the key components and enhancements in the proposed framework for Facial Beauty Prediction (FBP): 1- Shifted Patch Tokenization (SPT)

Enhance the model ability to capture local features by introducing overlaps between patches, mitigating the loss of spatial detail that can occur when dividing the image into non-overlapping patches. The input image is split into patches, but these patches are "shifted" to overlap partially with neighboring patches. This overlapping ensures that spatial continuity and local relationships are better preserved. Improved representation of fine-grained features, critical for tasks like facial beauty prediction where subtle details can influence the outcome. 2- Vision Transformer (ViT)

Token Embedding:
After tokenizing the patches, each patch is linearly projected to a fixed-dimensional embedding vector. Positional encodings are added to these embeddings to maintain spatial order.

Transformer Blocks:
Self-Attention Mechanism:
Captures global dependencies by weighing the importance of each patch relative to every other patch. Enables the model to learn relationships between distant regions of the face, such as the alignment of facial symmetry. Feed-Forward Neural Networks (FFNN):
Provides non-linear transformations for enhanced feature extraction. Skip Connections: Aid in stabilizing training and help the model learn efficiently in deeper networks.

Locality Self-Attention (LSA):
While traditional self-attention focuses on global relationships, LSA emphasizes local dependencies within a patch and its immediate neighbors. This concept is particularly suited for facial beauty tasks, as local regions (e.g., eyes, lips) are critical for attractiveness assessment.

Regression Token:
A special token is prepended to the input sequence, aggregating information across all patches to predict the final beauty score.

Additional Fully Connected Layers:
Two fully connected layers are added post-transformer blocks to refine the regression output. Together, these components address one of the ViT's primary challenges: its lack of inherent locality inductive bias compared to CNNs. By combining the overlapping patch tokenization (SPT) and a locally focused self-attention mechanism (LSA), the model achieves a balance between global and local feature learning. Figure 3.8 illustrates this framework, showing the flow from input images through shifted patch tokenization, processing in Vision Transformer blocks, and final score regression. This architecture highlights how ViTs can be adapted and

enhanced for specific tasks like Facial Beauty Prediction, setting a foundation for further research and practical applications.

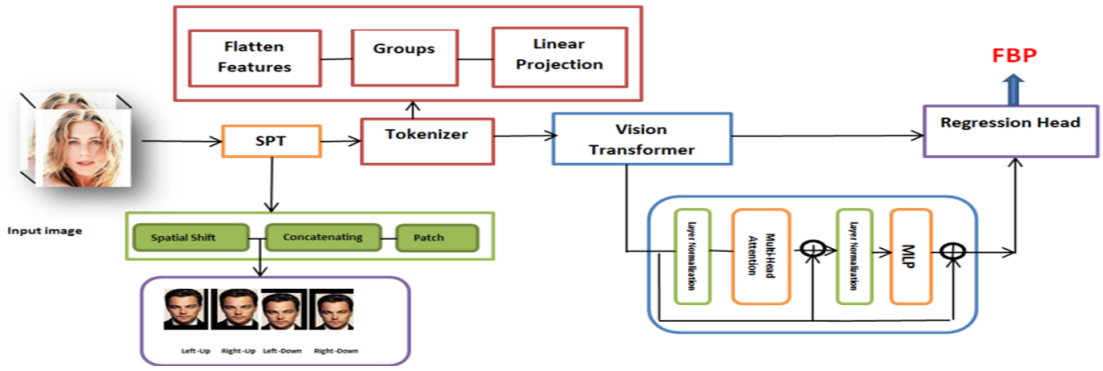


Figure 3.8: The proposed framework for regression of Facial Beauty Prediction

Shifted patch tokenization (SPT) [151] is a technique used in Vision Transformer (ViT) architectures to enhance their performance on small-sized datasets. ViTs are a type of deep learning model that utilizes transformer attention to process images. However, Vision Transformers (ViTs) have a limited receptive field, which implies that they can only perceive a small section of the image at a time. This can make it difficult for ViTs to learn from small-sized datasets, as there is not enough data for the model to learn from[152].

SPT addresses this problem by shifting the image in diagonal directions and then concatenating the shifted images with the original image. This expands the receptive field of the model, enabling it to capture more of the image simultaneously. As a result, Vision Transformers (ViTs) with Shifted Patch Tokenization (SPT) can achieve better performance on small-sized datasets than ViTs without SPT [153]. Here is a more detailed explanation of SPT:

- Start with an image.
- Shift the image in diagonal directions. This can be done by padding the image with zeros and then extracting subimages of the padded image.
- Concatenate the diagonally shifted images with the original image. This creates a new image that is larger than the original image.
- Extract patches of the concatenated images. This will create a set of tokens that represent the image.
- Flatten the spatial dimension of all patches. This will create a one-dimensional array of tokens.
- Layer normalize the flattened patches and then project it. This will transform the tokens into a format that can be used by the ViT.

Shifted Patch Tokenization

The SPT is a technique introduced to enhance the locality inductive bias of ViTs, enabling them to perform better with less data [154].

This allows ViTs trained with SPT to achieve better performance on smaller datasets compared to standard ViT architectures. The SPT concept is illustrated in Figure 3.7.

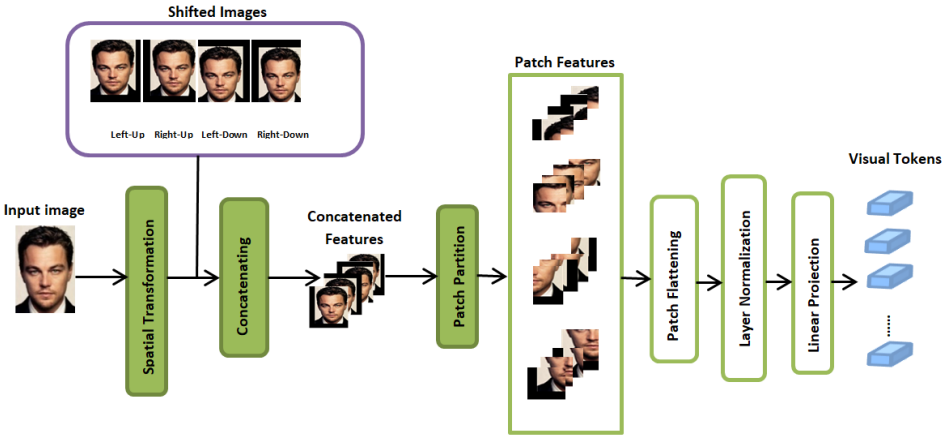


Figure 3.9: The architecture Shifted Patch Tokenization for Facial Beauty Prediction

1. Input Preparation: Take the original input image.
2. Shifted Image Creation: Shift the original image diagonally in four different directions:
   - Up-left by half the patch size.
   - Up-right by half the patch size.
   - Down-left by half the patch size.
   - Down-right by half the patch size.
3. Image Concatenation: Combine the original image with the four shifted versions into a single larger image. This creates an image with a larger size and includes overlapping information between neighboring patches.
4. Patch Extraction: Divide the concatenated image into patches of a predetermined size, similar to how the original image would be divided in a standard ViT.
5. Preprocessing: Apply any necessary preprocessing steps to the extracted patches, such as flattening the spatial dimensions and layer normalization.

6. Projection: Project the preprocessed patches using a linear transformation to embed them in a higher dimensional space. This step is crucial for extracting meaningful features from the patches.

7. Subsequent Processing: The projected patches are then fed into the downstream layers of the ViT architecture, such as the self-attention and feedforward layers, for further processing and learning relationships between the patches.

In the given context, let's consider an input image $x \in R^{H \times W \times C}$. , where H represents the height, W denotes the width, and C signifies the number of channels in the image.Then, the resulting image is divided into non-overlapping patches and the patches are flattened as in Eq. 1 [153].

$$P(x) = [x_p^1; x_p^2; ...; x_p^N] \tag{3.5}$$

where, $x_p^i \in R^{P^2 xC}$ is $i-th$ flattened vector. P is the patch size and $N = \frac{HW}{P^2}$ the number of patches.
After that, visual tokens (VT) are obtained through layer normalization (LN) and projected by a linear layer (LL). The entire procedure may be expressed as

$$VT(x) = LN(LL([x_p^1; x_p^2; ...; x_p^N])) \tag{3.6}$$

To use SPT as a Patch Embedding Layer, we supplemented the SPT output with a positional embedding variable. The entire procedure is designed as [153]:

$$VT_{pe}(x) = VT(x) + E_{pos} \tag{3.7}$$

where, E_{pos} is the learnable positional embedding and VTpe(x) is the ultimate output to be processed by the rest of the model.

Shifted Patch Tokenization (SPT) [151] has an impact on the computational complexity of Vision Transformers by introducing overlap between patches, which can affect the overall performance and efficiency of the model. Specifically:

- Computational Complexity: SPT introduces a more intricate process of patch extraction by creating overlap between adjacent patches through shifting strategies. This overlapping technique can potentially increase the computational complexity compared to traditional non-overlapping patch tokenization methods.

- Performance Improvement: Despite the potential increase in computational complexity, SPT aims to enhance the performance of Vision Transformers,

especially on small datasets, by introducing a locality inductive bias through shifted patch representations.

Locality Self-Attention Mechanism

Both LSA and SPT are innovative techniques designed to overcome challenges faced by Vision Transformers (ViTs) when dealing with small-size datasets. The locality self-attention mechanism is introduced to help the attention mechanism work more effectively on the dataset, as is presented in Figure.3. Before illustrating the innovation part, it is necessary to make clear why we need the refinement.the size of FBP image datasets is often small. Restricts information exchange in self-attention to local neighborhoods in the input data (often 2D feature maps). Instead of attending to every element, each element focuses on its nearby neighbors. The core of LSA [151] is the diagonal masking and learnable temperature scaling,

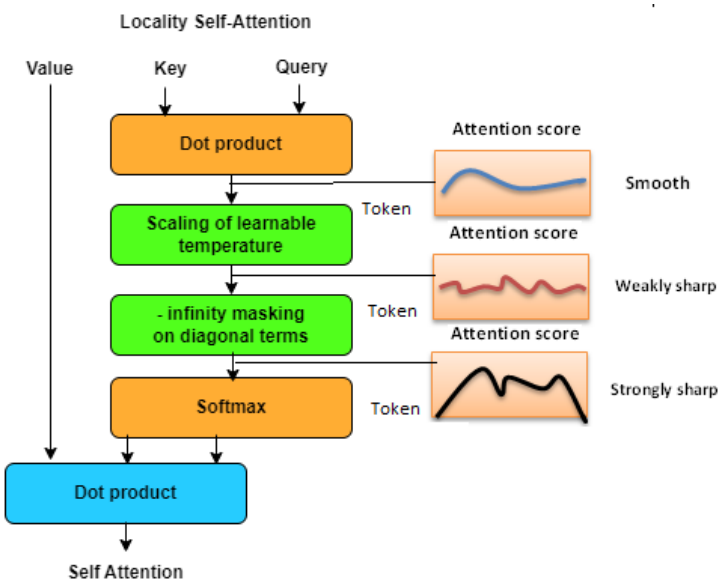


Figure 3.10: The architecture of Locality Self Attention

can be used in Locality Self-Attention (LSA) to control the "sharpness" of attention within local neighborhoods.understanding of the diagram in figure 3:

- Smooth: This likely represents standard LSA without any additional modifications. The attention scores are spread more evenly across neighboring tokens, indicating less focused attention.
- Weakly Sharp: This could represent using diagonal masking. By masking out attention scores beyond a certain diagonal, this restricts attention to closer neighbors, creating a "weaker" focus.

- Strongly Sharp: This could represent using learnable temperature scaling. Here, a learnable parameter ("temperature") is used to scale the attention scores, making them more peaked towards the highest scoring neighbors. This leads to a "stronger" and more focused attention.
- Token: This shows the position of each token in the sequence.
- Attention Score: This shows the attention score that each token receives from its neighbors (represented by the color intensity).

The diagram highlights how LSA can be fine-tuned for different levels of attention "sharpness" using diagonal masking and learnable temperature scaling. This allows the model to adapt its attention patterns to better capture relevant information based on the specific task and data [155].

Diagonal masking

The purpose of diagonal masking in Locally-Grouped Self-Attention (LSA) is to prevent the model from attending to itself, which can lead to overfitting. Diagonal masking is a binary mask that is applied to the attention matrix, setting the diagonal elements to zero. This ensures that the model does not attend to the current position, which can improve the model's ability to capture meaningful information from the input. Through diagonal masking, LSA's attention focuses more on meaningful regions of the feature map, which can improve the model's performance on various tasks. Diagonal masking is typically applied in conjunction with learnable temperature scaling, which allows the model to adjust the sharpness of the attention distribution, further improving its ability to focus on specific local regions of the input[156] .
Diagonal masking in Locally-Grouped Self-Attention (LSA) works by setting the diagonal elements of the attention matrix to zero. This mechanism prevents the model from attending to itself, which can lead to overfitting. The attention matrix is computed using the SoftMax function, which computes the attention distribution for each element in the input. The diagonal mask is applied to this attention matrix before computing the final attention weights. The equation for the attention matrix with diagonal masking can be represented as [151]:

$$Attention(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d_k}})V. \tag{3.8}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. The diagonal masking operation can be

represented as:

$$DiagonalMask(M) = \begin{cases} M_{ij} & \text{if i} \neq \text{j .} \\ 0 & \text{if i} = \text{j .} \end{cases} \tag{3.9}$$

The final attention matrix with diagonal masking is computed as:

$$AttentionWithDiagonalMask(Q, K, V) = DiagonalMask(SoftMax(\frac{QK^T}{\sqrt{d_k}}))V. \tag{3.10}$$

Diagonal masking in LSA prevents the model from attending to itself, which can lead to overfitting. This mechanism improves the model's ability to capture meaningful information from the input, and it is typically applied in conjunction with learnable temperature scaling to further improve the model's attention mechanism Diagonal masking in Locally-Grouped Self-Attention (LSA) affects the self-attention mechanism by preventing the model from attending to itself, which can lead to overfitting. The diagonal mask is a binary mask that is applied to the attention matrix, setting the diagonal elements to zero. This ensures that the model does not attend to the current position, which can improve the model's ability to capture meaningful information from the input. By preventing the model from attending to itself, diagonal masking improves the model's ability to capture meaningful information from the input, and it is typically applied in conjunction with learnable temperature scaling to further improve the model's attention mechanism. This mechanism improves the model's ability to focus more on meaningful regions of the feature map, which can improve the model's performance on various tasks[153].

learnable temperature scaling

The purpose of learnable temperature scaling in Locally-Grouped Self-Attention (LSA) is to adjust the sharpness of the attention distribution, allowing the model to focus on more or fewer elements in the input. This mechanism enables LSA to choose the appropriate level of concentration for the attention distribution, which can be beneficial in various scenarios. Learnable temperature scaling is applied to the SoftMax function, which is used to compute the attention distribution. The SoftMax function can regulate the output distribution's smoothness by temperature scaling. In LSA, the temperature parameter is learned during training, which allows the model to adapt to the specific requirements of the task at hand. By learning the temperature parameters, LSA can sharpen the distribution of attention scores, which can improve the model's performance in tasks where focusing on specific local regions is crucial. This mechanism is particularly useful in vision

transformers, where LSA has been shown to improve performance on small-size datasets.

In summary, learnable temperature scaling in LSA allows the model to adjust the sharpness of the attention distribution, which can improve the model's ability to focus on specific local regions of the input. This mechanism is essential for LSA's ability to balance the trade-off between locality and global information in the input[151]. Learnable temperature scaling in Locally-Grouped Self-Attention (LSA) affects the output distribution by sharpening the attention distribution. This mechanism enables the model to focus more on specific elements in the input, thereby enhancing its ability to capture local information. By adjusting the sharpness of the attention distribution, learnable temperature scaling allows LSA to adapt to the specific requirements of the task at hand. This can be particularly beneficial in scenarios where focusing on local regions of the input is crucial, such as in vision transformers. Therefore, learnable temperature scaling plays a key role in balancing the trade-off between locality and global information in the input, ultimately influencing the model's attention mechanism and its ability to process input data effectively[151]. The equation for learnable temperature scaling in Locally-Grouped Self-Attention (LSA) is not explicitly provided in the search results, but the concept is described in terms of adjusting the SoftMax function's temperature parameter. The SoftMax function is used to compute the attention distribution, and its temperature parameter is typically set to 1. In LSA, this temperature parameter is learned during training, which allows the model to adapt to the specific requirements of the task at hand. The SoftMax function with learnable temperature scaling can be represented as[153]:

$$SoftMax(x_i) = \frac{e^{x_i/\tau}}{\Sigma_{j=1}^{N} e^{x_j/\tau}}. \tag{3.11}$$

where, x_i is the input to the SoftMax function, N is the number of elements in the input, and τ is the learnable temperature parameter. In LSA, the temperature parameter τ is learned during training, which allows the model to adjust the sharpness of the attention distribution, thereby improving its ability to focus on specific local regions of the input

The key findings of the study on facial beauty prediction using Vision Transformers include:

- STP-LSA-ViT-FBP Model: The study introduces the STP-LSA-ViT-FBP model, which is a deep learning model that utilizes Vision Transformers to predict facial beauty. This model is designed to address the challenges of predicting facial beauty, especially with small-sized datasets.

- Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA): The STP-LSA-ViT-FBP model incorporates innovative techniques like Shifted Patch Tokenization and Locality Self-Attention. These techniques help overcome the lack of locality inductive bias and enable the system to learn effectively even with limited data.

- Accuracy and Interpretability: The study highlights that the STP-LSA-ViT-FBP model is more accurate and interpretable compared to previous methods used for facial beauty prediction. This indicates a significant improvement in the efficiency and performance of predicting facial beauty.

- Potential Applications: The research suggests that the STP-LSA-ViT-FBP model has the potential to be utilized in various applications related to facial beauty prediction. This includes applications in fields such as cosmetic surgery, beauty analysis, and facial recognition technology.

- Comparison with Convolutional Neural Networks (CNNs): Vision Transformers have shown comparable or even superior results to state-of-the-art deep convolutional neural network techniques in predicting facial beauty. This indicates the promising performance of Vision Transformers in this domain.

These findings underscore the potential of Vision Transformers in revolutionizing the prediction of facial beauty, offering more accurate and interpretable results that can be applied in diverse real-world scenarios.

The application of Vision Transformers in predicting facial beauty has significant implications for real-life scenarios in the beauty industry and healthcare. Here are some ways this technology can be applied:

1. Beauty Industry:

    - Virtual Try-On: Beauty brands can use Vision Transformers to create virtual try-on experiences for customers, allowing them to see how different makeup looks or skincare products would appear on their face before making a purchase.
    - Personalized Recommendations: By analyzing facial features and beauty preferences, Vision Transformers can provide personalized product recommendations tailored to individual customers, enhancing the shopping experience and increasing customer satisfaction.
    - Beauty Analysis: Beauty salons and cosmetic clinics can utilize Vision Transformers to analyze facial features and recommend personalized beauty treatments or procedures based on individual characteristics, leading to more targeted and effective beauty solutions.

2. Healthcare:

    - Facial Reconstruction: In healthcare, Vision Transformers can assist in facial reconstruction procedures for patients who have undergone trauma or surgery, helping reconstruct facial features with precision and accuracy.
    - Dermatological Diagnosis: Vision Transformers can aid dermatologists in diagnosing skin conditions and analyzing facial features for early detection of skin diseases or abnormalities, enabling timely treatment and care.
    - Plastic Surgery Planning: Surgeons can use Vision Transformers to simulate and plan plastic surgery procedures, allowing them to visualize potential outcomes and optimize surgical plans for better results and patient satisfaction.

3. Facial Recognition Technology:

    - Security and Access Control: Vision Transformers can enhance facial recognition systems used for security purposes, access control, and identity verification, improving the accuracy and efficiency of facial recognition technology in various industries.
    - Emotion Analysis: By analyzing facial expressions and features, Vision Transformers can be applied in healthcare settings to assess patients' emotional states, monitor mental health conditions, and provide personalized emotional support and care.

Overall, the application of Vision Transformers in the beauty industry and healthcare sectors can revolutionize how facial beauty is predicted, analyzed, and utilized in various practical applications, leading to enhanced customer experiences, personalized treatments, and improved outcomes in both industries.

3.4 Conclusion

In this chapter, we explored how deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have been employed to analyze facial features and predict facial beauty ratings with remarkable accuracy. These advanced methodologies have addressed several challenges in the domain, including the scarcity of high-quality datasets and the inherently subjective nature of beauty assessments.

By utilizing CNNs, which are effective at capturing hierarchical spatial features, and ViTs, which excel at modeling global dependencies across image patches, we have developed models that improve upon traditional methods. These advancements not only enhance the precision of beauty prediction but also provide more consistent results that can be interpreted across various contexts. As a result, such deep learning-based models are becoming increasingly valuable tools for industries like cosmetics, fashion, and entertainment, where accurate assessments of facial aesthetics can drive product development, marketing strategies, and user personalization.

In the next section, we will present the experimental results that validate the effectiveness of the proposed methods. This includes detailed discussions on model performance metrics, comparisons to baseline methods, and insights into how the proposed architectures have contributed to improved prediction accuracy. The following analyses will focus on: A direct comparison of the proposed methods (e.g., Ensemble CNNs and ViT-FBP) against standard baseline models to highlight improvements in facial beauty prediction accuracy. Presentation of key performance indicators such as Pearson Correlation Coefficient (PCC), Mean Absolute Error (MAE), and other relevant statistical measures that demonstrate the robustness and reliability of the developed models. Visual demonstrations of model predictions, showcasing how the models analyze different facial features and provide ratings aligned with human judgment. Analysis of how well the models generalize to unseen data and how they handle diverse faces across various demographics (age, gender, ethnicity).

This empirical evaluation will provide clear evidence of the contribution of deep learning models to the field of facial beauty prediction, setting the stage for potential real-world applications.

Chapter 4

Results and discussion

4.1 Introduction

Facial beauty prediction is a subcategory of facial recognition that utilizes advanced machine learning algorithms to determine the attractiveness of a person's facial features [157]. The process is conducted by experts in the field of facial recognition who have the necessary training to interpret the data collected from the AI algorithms.
After the study the convolutional neural networks and vision Transformers applied on predictions of the facial beauty (see chapter 3). We carry out in this chapter a detailed study of the results obtained by the algorithms applied on the facial beauty prediction. A comparative study with current facial beauty prediction algorithms such as PI-CNN, CNN with SCA, CNN + LDL, ResNet-18 based AaNet, R3CNN and CNN-ER is thus carried out.

4.2 Performance Evaluation

Depending on the application sought, the beauty prediction algorithm evaluated must be able to verify a certain number of prediction quality criteria, among which we can cite the mean absolute error (MAE), the error root mean square (RMSE), and the Pearson correlation (PC) [158]. At this stage and after having obtained our prediction results, we must evaluate the performance of our system from the test data. And since the model belongs to regression problems, we will apply the specified metrics to this type of problem [159]. The metrics used are:

4.2.1 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a regression metric that measures the average magnitude of the errors between predicted values and actual values. It is calculated

by taking the average of the absolute values of the errors [160]. MAE is a scale-dependent metric, meaning that it is sensitive to the scale of the data. For example, if the data is in a different scale, the MAE will also be different. This can make it difficult to compare MAE values across different datasets.
However, MAE is a simple and easy-to-understand metric. It is also robust to outliers, making it a good choice for regression tasks with noisy data [161]. MAE is typically used to evaluate the performance of regression models. It can also be used to compare the performance of different regression models on the same dataset.
The mean absolute error (MAE) is defined by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i| \tag{4.1}$$

Where y_i is the actual value,$\widehat{y}_i$ is the predicted value and N the number of all predicted values.
MAE is a useful metric for evaluating the performance of regression models, but it is important to keep its limitations in mind. It is a scale-dependent metric, and it is not sensitive to the direction of the errors[152]. MAE is a versatile metric that can be used to evaluate the performance of regression models in a variety of different domains.

4.2.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a regression metric that measures the standard deviation of the residuals between predicted values and actual values. It is calculated by taking the square root of the mean of the squared residuals[160]. RMSE is a scale-invariant metric, meaning that it is not sensitive to the scale of the data. This makes it easy to compare RMSE values across different datasets. RMSE is also sensitive to the direction of the errors, meaning that it penalizes large errors more heavily than small errors. This makes it a good choice for regression tasks where large errors are more costly [161]. RMSE is typically used to evaluate the performance of regression models. It can also be used to compare the performance of different regression models on the same dataset.
Here is the formula for RMSE:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i|^2} \tag{4.2}$$

Where y_i is the actual value,$\widehat{y}_i$ is the predicted value and N the number of all predicted values.
RMSE is a useful metric for evaluating the performance of regression models, but it is important to keep its limitations in mind. It is not robust to outliers, meaning that it can be skewed by a small number of very large errors [162]. Overall, RMSE is a widely used and well-understood regression metric. It is a good choice for regression tasks where large errors are more costly and the data is not noisy.

4.2.3 Pearson correlation

Pearson correlation is a statistical measure that quantifies the linear relationship between two variables. It is calculated by taking the covariance of the two variables and dividing it by the product of their standard deviations [159].
Pearson correlation is a scale-invariant metric, meaning that it is not sensitive to the scale of the data. It is also a robust metric, meaning that it is not affected by outliers. Pearson correlation is typically used to measure the strength and direction of the relationship between two continuous variables. It can also be used to compare the relationships between two or more variables.
Here is the formula for Pearson correlation [163]:

$$PC = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\widehat{y}_i - \bar{\widehat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\widehat{y}_i - \bar{\widehat{y}})^2}}, \tag{4.3}$$

Where y_i is the actual value,$\widehat{y}_i$ is the predicted value and N the number of all predicted values.
Pearson correlation coefficients can range from -1 to 1. A coefficient of -1 indicates a perfect negative correlation, a coefficient of 1 indicates a perfect positive correlation, and a coefficient of 0 indicates no correlation [163].
Pearson correlation is a versatile metric that can be used to measure the strength and direction of the relationships between two or more variables in a variety of different domains.

4.3 Optimization Method

In machine learning, the word "optimization" is frequently used since the goal is to discover the least amount of error between the actual output and the predicted output. Thus, an optimizer is a crucial component, which is described as an algorithm that instructs us on how to arrive at the minimum. For example Gradient Decent and Stochastic Gradient Decent are two types of optimization methods. This sec-

tion describes a specific sort of stochastic gradient descent method and its extension.

4.3.1 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent (SGD) [164] is an iterative optimization method commonly used in machine learning to optimize an objective function with suitable smoothness properties. It is a stochastic approximation of gradient descent optimization, where instead of calculating the gradient from the entire dataset, SGD estimates the gradient from a single training example or a small batch of examples. This randomness in selecting data points reduces computational inefficiency, making SGD computationally efficient, especially when dealing with large datasets. By updating the model parameters based on random training examples, SGD offers advantages such as speed, memory efficiency, and the ability to escape local minima. However, it can lead to noisy updates and slower convergence compared to other gradient descent variants like batch gradient descent.

4.3.2 Adam

Adam [147] is an adaptive gradient descent algorithm. It stands for Adaptive Moment Estimation. Adam is a popular algorithm for training machine learning models, including Transformers. It is a modified version of the stochastic gradient descent (SGD) algorithm that adapts the learning rate for each parameter based on the first and second moments of the gradients [147].
Adam is a very efficient algorithm and it is able to converge to the optimal solution quickly. It is also very robust to noise in the data. Here is a brief overview of how Adam works:

- Initialize the parameters of the model and the learning rate.
- Compute the gradients of the loss function with respect to the parameters.
- Compute the first and second moments of the gradients.
- Update the parameters using the Adam update rule.
- Repeat steps 2-4 until the model converges.

Adam is a powerful algorithm that can be used to train a wide variety of machine learning models. It is particularly well-suited for training Transformers, which are complex models with many parameters. Here are some of the benefits of using Adam:

- It is very efficient and can converge to the optimal solution quickly.
- It is very robust to noise in the data.
- It is easy to implement and tune.

Adam is a popular choice for training Transformers and other machine learning models because it is efficient, robust, and easy to use.

4.3.3 AdamW

The AdamW optimization method is an extension of the Adam algorithm, designed to improve training speeds in deep neural networks by modifying the typical implementation of weight decay [165]. In AdamW, weight decay is decoupled from the gradient update, unlike in traditional Adam where L2 regularization is usually implemented with a modification that includes weight decay in the gradient update. This modification allows AdamW to adjust the weight decay term separately in the gradient update, leading to better generalization and improved model performance compared to standard Adam. By decoupling weight decay from the gradient update, AdamW addresses issues related to L2 regularization and provides models that generalize more effectively.

The difference between the Adam and AdamW optimization methods lies in how they handle weight decay during the optimization process:

- Adam: In the standard Adam optimization algorithm, weight decay (L2 regularization) is usually implemented by adding the weight decay term to the gradients during the update step. This means that weight decay is included in the gradient calculation, affecting how the weights are updated.
- AdamW: On the other hand, AdamW decouples weight decay from the gradient update. In AdamW, the weight decay term is adjusted separately after controlling the parameter. This adjustment ensures that the weight decay does not end up in the moving averages and is only proportional to the weight, leading to better generalization and improved model performance compared to standard Adam.

While both Adam and AdamW are adaptive optimization algorithms, AdamW specifically addresses the issue of weight decay implementation by adjusting the weight decay term separately from the gradient update, resulting in models that generalize better and compete with stochastic gradient descent (SGD) more effectively.

4.4 Development tools and language

We realized our program using: Colaboratory: ("Colab" for short) is a data analysis and machine learning tool that lets combine executable Python code and rich text with graphics, images, HTML, LaTeX and more in a single document stored in Google Drive. It connects to powerful Google Cloud Platform runtimes.

Figure 4.1: Google Colab logo

Google Colab is a free Jupyter notebook that allows running Python in the browser without the need for complex configuration. It comes with Python installed and all major Python libraries installed. It is also integrated with free GPUs. Python is an ideal language for writing scripts in a very simple way and the rapid development of applications in many fields and on most platforms, all thanks to its many libraries.

Figure 4.2: Notebook Jupyter and Python

1. Libraries

   In our program, we used a set of libraries to use predefined functions. They

are called at the beginning of the program.

2. Keras
   Keras is an open source library written in python that runs on the machine learning platform TensorFlow and also it is easy to use for developing and evaluating deep learning models because it allows creating layers very easily for neural networks or to set up complex architectures.

3. TensorFlow
   TensorFlow is an open source machine learning library, created by Google (the second generation of the Google Brain system) for developing and running machine and deep learning applications. It is a toolkit for solving extremely complex mathematical problems and high performance numerical computing with ease. Its flexible architecture makes it easy to deploy compute on a variety of platforms (CPU, GPU and TPU). Scikit-learn "Sklearn": is an important library of tools dedicated to machine learning and data science in the Python universe.

4. Numpy
   The term NumPy is actually short for Numerical Python. It is an Open Source library in Python language. This tool is used for scientific programming in Python, and in particular for data science programming knowing that it offers a large number of routines for quick access to this data. It is intended to handle matrices or multidimensional arrays (array) as well as mathematical functions operating on these arrays. Matplotlib: A Python library, intended for plotting and visualizing data in the form of graphs. It can be combined with the scientific computing python libraries NumPy and SciPy and is notably used on web application servers, shells and Python scripts.

5. Pandas
   The name Pandas is actually a contraction of the term Panel Data, designating data sets that include observations over multiple time periods. The Pandas open-source software library is specifically designed for data manipulation and analysis in the Python language, in particular it offers data structures and operations for manipulating numerical arrays and time series to make working with data easier.

4.5 EXPERIMENTS

This section describes the experiments and assessment findings from the models algorithms used in this study. A summary of the training and test dataset is included in this section as well. This section also includes the findings from the proposed algorithms of DCNNs, including Mean models, various conventional pretrained models, and cutting-edge face beauty prediction categorization techniques. In this section, comparisons between the proposed algorithms of DCNNs and existing CNN models created from scratch are shown. Utilized resources all the experiments are implemented using TensorFlow, Keras API, and utilized Python programming in Google Colaboratory or CoLab. After uploading the dataset to Google Drive, we use Tesla GPU in the CoLab to execute our experiment. The deep learning approach has studied in earlier studies in classification problems of face beauty of different datasets. We thus established the identical hyper-parameters for all four deep learning models on the provided dataset to assess the models' performances while taking into account their impacts.

4.5.1 The dataset used

For our work, we used the SCUT-FBP5500 dataset.

The SCUT-FBP5500 data refers to a dataset specifically designed for facial beauty prediction (FBP) tasks. Here's a summary of its key aspects[166]:

Purpose:

- To train and evaluate multi-paradigm FBP models, allowing exploration of different approaches like appearance-based, shape-based, classification, regression, and ranking.
- To address limitations of existing FBP datasets by offering more diversity in:
  - Ethnicity: Includes Asian and Caucasian individuals.
  - Gender: Includes male and female individuals.
  - Age: Covers a range from 15 to 60 years old.
  - Beauty Scores: Provides subjective human-rated beauty scores ranging from 1 (least beautiful) to 5 (most beautiful).

Data Composition:

- Size: 5,500 frontal face images.
- Subsets: Divided into four equal subsets based on ethnicity and gender:

- 2,000 Asian females (AF)
- 2,000 Asian males (AM)
- 750 Caucasian females (CF)
- 750 Caucasian males (CM)

Data Labels:

Each image comes with two types of labels:

- Beauty Score: Assigned by human evaluators on a scale of 1 to 5.
- 86 Facial Landmarks: Markings on key facial features like eyes, nose, mouth, and eyebrows.

As shown in Figure 4.3, Female Asian samples and the corresponding scores are from right to left : (1.56; 2.45; 3.51; 4.28), Male Asian samples and the corresponding scores are from right to left : (1.53; 2.46; 3.53; 4.23), Female Caucasian samples and the corresponding scores are from right to left : (1.93; 2.45; 3.66; 4.45) and Male Caucasian samples and the corresponding scores are from right to left : (1.53; 2.66 ; 3.45; 4.2)[167].

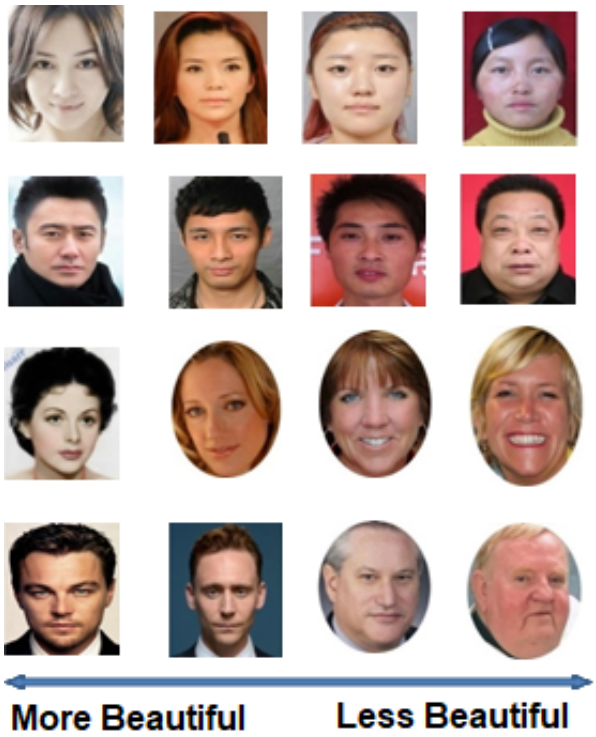


Figure 4.3: Images of various facial features and beauty ratings from the SCUT-FBP5500 benchmark dataset.

4.5.2 The proposed approach EN-CNN

The proposed EN-CNN (Ensemble of Deep Convolutional Neural Networks) model is designed to predict facial beauty scores using an ensemble regression approach. The ensemble consists of four pre-trained deep convolutional neural networks (DenseNet201, InceptionV3, MobileNetV2, and EfficientNetB7), each fine-tuned for the specific task of facial beauty prediction. Each model contributes to the final prediction combining their outputs for robust and accurate beauty score estimation.To evaluate the performance, the SCUT-FBP5500 dataset is partitioned into 80% training data and 20% testing data, ensuring sufficient samples for both model training and validation. This setup ensures that the models generalize well to unseen data, a critical factor in reliable attractiveness estimation.

The performance comparison of the proposed EN-CNN model with existing state-of-the-art methods for facial beauty prediction on the SCUT-FBP5500 dataset is presented in Table 4.1. The Mean Absolute Error (MAE) of the EN-CNN model is 0.1783, the lowest among all compared methods, indicating that our model makes smaller average errors in beauty score predictions. Similarly, the Root Mean Square Error (RMSE) of 0.2564 demonstrates the robustness of the EN-CNN in minimizing larger errors compared to other approaches. These results reflect the enhanced generalization and fine-grained prediction capabilities of the EN-CNN model. The Pearson Correlation (PC) score of 0.9469 achieved by EN-CNN is significantly higher than all baseline methods. This suggests that the EN-CNN model aligns more closely with human assessments of facial beauty. Compared to the closest competitor, Vahdati et al.'s model (PC = 0.9372), EN-CNN shows a notable improvement, showcasing the effectiveness of the ensemble and the carefully tuned architecture.Comparison with Other Models: AlexNet, ResNet-18, and ResNeXt-50: While these standard architectures perform reasonably well, their MAE and RMSE scores are significantly higher, and their PC values are lower compared to EN-CNN. This indicates that single models pre-trained on ImageNet lack the specialized capacity to predict facial beauty accurately. CNN-SCA, PI-CNN, and CNN + LDL: These models utilize additional techniques like label distribution learning or spatial attention mechanisms, resulting in improved performance over standard architectures. However, they still fall short compared to the ensemble approach used in EN-CNN. NAS4FBP and Vahdati et al.: Both models are among the strongest competitors, leveraging neural architecture search (NAS) or fine-tuning with VGGFace2 datasets. Despite their strengths, EN-CNN surpasses these models by combining diverse CNNs to extract complementary features.

Table 4.1: Performance comparisons EN-CNN on the SCUT-FBP5500 dataset

Methods	Pre-training	MAE ↓	RMSE ↓	PC ↑
AlexNet [80]	ImageNet	0.2651	0.3481	0.8634
ResNet-18 [81]	ImageNet	0.2419	0.3166	0.8900
ResNeXt-50 [81]	ImageNet	0.2291	0.3017	0.8997
CNN – SCA [71]	ImageNet	0.2287	0.3014	0.9003
PI-CNN [78]	ImageNet	0.2267	0.3016	0.8978
CNN + LDL [63]	ImageNet	0.2201	0.294	0.9031
ResNet-18 based AaNet [166]	ImageNet	0.2236	0.2954	0.9055
R3CNN [76]	ImageNet	0.2120	0.2800	0.9142
CNN-ER [63]	VGGFace2	0.2009	0.2650	0.9250
NAS4FBP Net [168]	ImageNet	0.1939	0.2579	0.9275
Vahdati at al. [166]	VGGFace2	0.1833	0.2422	0.9372
EN-CNN Ours	**ImageNet**	**0.1783**	**0.2564**	**0.9469**

The proposed EN-CNN architecture comprises 45.33 million parameters, strategically distributed across its constituent models: DenseNet201: 19.30 million parameters, contributing strong feature propagation and reuse. InceptionV3: 22.85 million parameters, capturing multi-scale features effectively with its inception modules. MobileNetV2: 2.91 million parameters, offering lightweight computations optimized for mobile and edge devices. EfficientNetB7: 3.17 million parameters, leveraging compound scaling to improve efficiency. This distribution ensures that each model brings unique strengths to the ensemble while maintaining computational feasibility. CNN-SCA (6.75M parameters): Despite its lightweight nature, CNN-SCA underperforms (MAE = 0.2287) compared to EN-CNN, highlighting the advantage of using a diversified ensemble approach over a single architecture. ResNeXt-50 (25.03M parameters): While larger than CNN-SCA, ResNeXt-50 achieves better performance (MAE = 0.2291) but falls short of EN-CNN, indicating that parameter count alone does not guarantee superior performance. AlexNet (62.38M parameters): Despite having more parameters, AlexNet achieves lower performance (MAE = 0.2651), demonstrating the importance of architecture optimization over parameter quantity. Vahdati et al.: Although highly optimized with 36.67 million parameters, Vahdati et al.'s model achieves a Pearson Correlation (PC) of 0.9372, which is surpassed by EN-CNN's PC of 0.9469.

EN-CNN leverages 45.33M parameters effectively, achieving: The lowest MAE (0.1783) among all compared methods. The highest PC (0.9469), showing strong alignment with human assessments. A balanced computational load by combin-

ing lightweight (e.g., MobileNetV2, EfficientNetB7) and high-capacity models (DenseNet201, InceptionV3). This design ensures that EN-CNN is not only powerful but also efficient, making it suitable for real-world applications. Figure 4.4 illustrates the comparison of expected scores.

Figure 4.4: Comparisons of the ground-truth, and predicted scores given by EN-CNNs

The connection between the forecast and the actual data is illustrated in Figure 4.5. This model calculates its parameters using data from numerous facial beauty scores that fall within a specified range because our goal is to predict facial beauty scores. The ground truth corresponds to the greatest number of forecast values; this may be inferred.

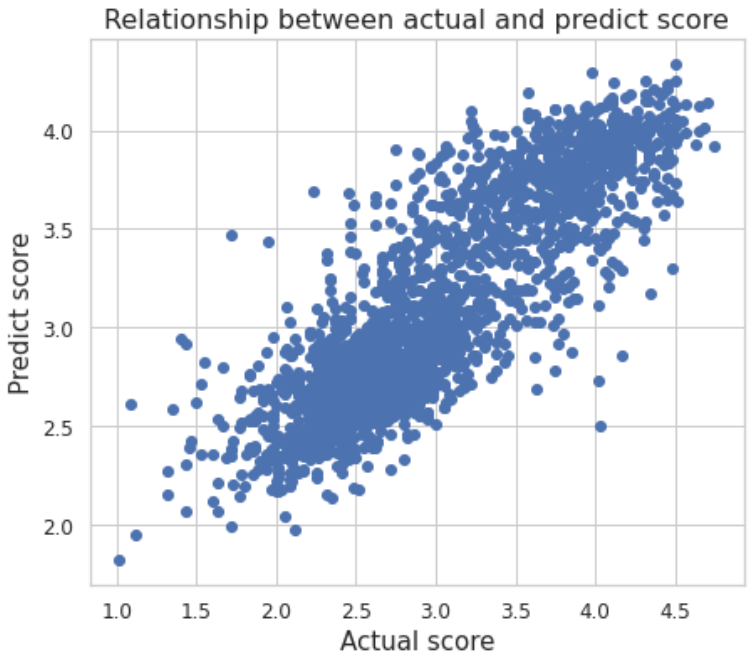


Figure 4.5: The relationship between a ground-truth and prediction.

Saliency and heat maps were created for each image to analyze the locations of faces. Saliency maps offer a visual representation of the degree to which particular points in an input image stand out more than their adjacent points. Heat maps display all regions of the image along with each facial landmark point. The figure

4.6 displays a saliency map and heatmap for several faces in the test set.

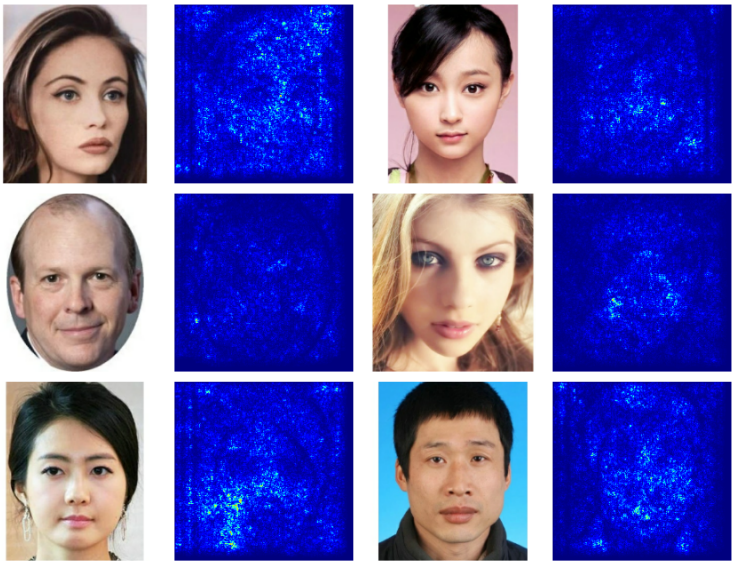


Figure 4.6: Saliency map (left) and heat maps (right) for face beauty

The proposed EN-CNN model demonstrates state-of-the-art performance in facial beauty prediction, achieving the highest correlation with human judgment (PC = 0.9469) and the lowest error rates (MAE = 0.1783, RMSE = 0.2564). These results validate the effectiveness of the ensemble strategy and the careful architectural design. EN-CNN sets a new benchmark for facial beauty prediction on the SCUT-FBP5500 dataset and underscores the potential of ensemble deep learning models in solving subjective tasks.

4.5.3 The proposed approach E-CNN

To validate the effectiveness of the E-CNN model, experiments were conducted using the SCUT-FBP5500 dataset. The dataset was split into 80% for training and 20% for testing, and the performance was compared against several state-of-the-art methods, as summarized in Table 4.2. The proposed E-CNN model demonstrates superior performance, achieving: MAE: 0.1933 (lowest among all methods. RMSE: 0.2482 (indicating excellent error minimization).PC: 0.9350 (highest correlation with ground truth).

Table 4.2: Performance comparisons EN-CNN on the SCUT-FBP5500 dataset

Methods	Pre-training	MAE ↓	RMSE ↓	PC ↑
AlexNet [80]	ImageNet	0.2651	0.3481	0.8634
ResNet-18 [81]	ImageNet	0.2419	0.3166	0.8900
ResNeXt-50 [81]	ImageNet	0.2291	0.3017	0.8997
CNN – SCA [71]	ImageNet	0.2287	0.3014	0.9003
PI-CNN [78]	ImageNet	0.2267	0.3016	0.8978
CNN + LDL [63]	ImageNet	0.2201	0.294	0.9031
ResNet-18 based AaNet [166]	ImageNet	0.2236	0.2954	0.9055
R3CNN [76]	ImageNet	0.2120	0.2800	0.9142
CNN-ER [63]	VGGFace2	0.2009	0.2650	0.9250
NAS4FBP Net [168]	ImageNet	0.1939	0.2579	0.9275
E-CNN Ours	**ImageNet**	**0.1933**	**0.2482**	**0.9350**

The performance of our ensemble of deep convolutional neural networks (E-CNN) for facial beauty prediction is compared with various state-of-the-art models in terms of three key metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson Correlation (PC). These metrics are evaluated under 5-fold cross-validation, providing a comprehensive view of each model's predictive accuracy and robustness. We observe that earlier models, such as AlexNet and ResNet-18, which were pre-trained on the ImageNet dataset, exhibit relatively higher MAE and RMSE values, along with lower PC scores. AlexNet, for instance, achieves an MAE of 0.2651 and an RMSE of 0.3481, with a Pearson Correlation of 0.8634, indicating modest performance. ResNet-18 shows some improvement with an MAE of 0.2419 and a corresponding PC of 0.8900. Similarly, ResNeXt-50 demonstrates further gains, reducing the MAE to 0.2291, RMSE to 0.3017, and increasing the PC to 0.8997. These results highlight that although ImageNet pre-training provides a strong starting point, the architecture of the model plays a critical role in its ability to capture subtle facial beauty features. The progression from AlexNet to ResNeXt-50 shows the benefits of deeper and more advanced network architectures. The specialized models, such as CNN-SCA, R3CNN, and NAS4FBP Net, offer significant performance improvements, particularly in terms of reducing MAE and RMSE while enhancing Pearson Correlation. For example, CNN-SCA achieves an MAE of 0.2287 and a PC of 0.9003, which is a marked improvement over the earlier, more general models. R3CNN further reduces the MAE to 0.2120 and increases the PC to 0.9142, demonstrating the effectiveness of using more complex architectures specifically designed for facial beauty assessment. Notably,

the CNN-ER and NAS4FBP Net models, pre-trained on VGGFace2 and ImageNet, outperform their counterparts with MAEs of 0.2009 and 0.1939, respectively. Their higher Pearson Correlation scores of 0.9250 and 0.9275 suggest these models better capture the subjective and nuanced nature of facial beauty, likely due to their specialized architectures and enhanced feature extraction capabilities.

Performance of the Proposed Ensemble Model

Our proposed ensemble model (E-CNN), which integrates InceptionV3, MobileNetV2, and a custom CNN-based architecture, achieves the best performance among all models. With an MAE of 0.1933, an RMSE of 0.2482, and a Pearson Correlation of 0.9350, E-CNN surpasses both state-of-the-art deep networks and specialized models for facial beauty prediction. This improvement can be attributed to the ensemble approach, which leverages the strengths of multiple architectures to provide a more robust and generalizable model. The reduced MAE and RMSE reflect the model's enhanced precision in predicting beauty scores, while the high Pearson Correlation indicates a strong alignment with human-perceived beauty judgments. These results underscore the efficacy of combining different architectural representations, each contributing unique features that improve overall performance. The superior performance of our ensemble model highlights several key findings. First, the use of pre-trained models on large-scale datasets such as ImageNet provides a solid foundation, but further improvements are possible when these models are fine-tuned or enhanced with architectures specifically designed for facial analysis. Second, the ensemble approach significantly improves predictive accuracy by mitigating the limitations of individual models. By capturing a broader range of features, the ensemble model offers a more nuanced and reliable prediction of facial beauty.

A graph illustrating the loss function against the number of epochs is displayed in Figure 4.6. At the beginning of the training process, both the training and validation data are unstable. Optimal results in terms of time and effort are achieved with 600 training epochs. Continuing the training process will lead to overfitting the training data, reducing the accuracy and stability of the validation data.

In addition, we analyze the distribution of predictors from the network, as illustrated in Figure 4.7. The curve illustrates the estimation of prediction scores. It can be seen that there is a shift in the mean value of the dataset predictions. However, it can be noted that the prediction values follow the general behavior of the ground-truth distribution.

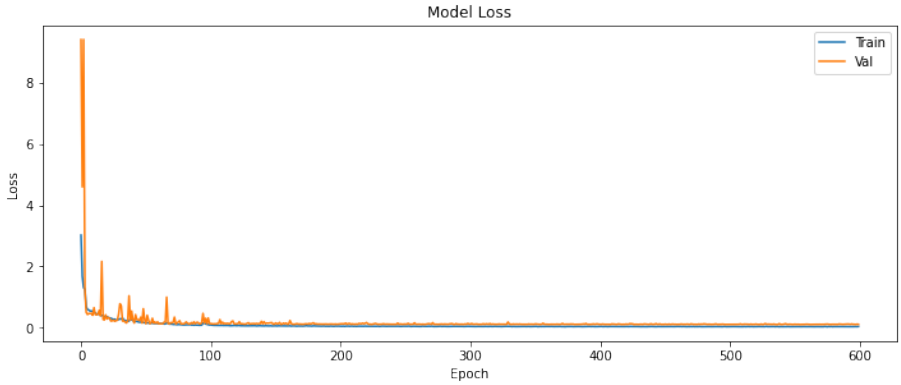


Figure 4.7: A graph of the loss function against the number of epochs. The blue curve is associated with training data loss and the red curve shows validation data loss

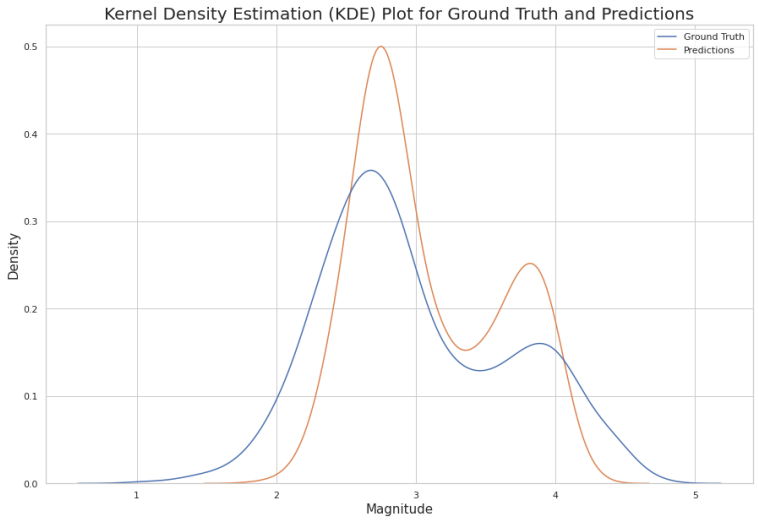


Figure 4.8: The blue curve is the ranks of ground-truth. The red curve is the ranks of prediction.

Figure 4.8 depicts the relationship between the ground truth and the prediction. Since our goal is to predict facial beauty scores, this model estimates its parameters using data from scores that fall within a certain range. It can be deduced that the ground truth correlates most with the predicted values.

Our ensemble-based approach demonstrates a significant advancement in facial beauty prediction, outperforming existing methods both in terms of error metrics and correlation with human judgments. This suggests that ensemble learning is a promising direction for improving the accuracy and reliability of models in subjective tasks like beauty prediction.

The results highlight the strength of the E-CNN approach, particularly its ability

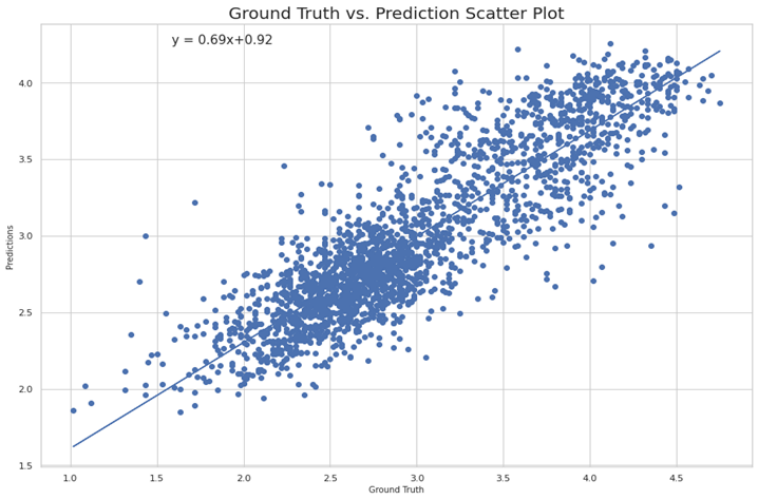


Figure 4.9: The relationship between a ground-truth and prediction.

to combine features from multiple architectures effectively. By integrating transfer learning, custom loss functions, and a robust ensemble method, the E-CNN achieves unparalleled accuracy in predicting facial beauty scores. This advancement demonstrates significant potential for applications in various domains, including cosmetics, entertainment, and fashion.

4.5.4 The proposed approach ViT-FBP

In this section, we introduce our proposed ViT-FBP architecture for facial beauty prediction, which leverages the strengths of Vision Transformers (ViTs). This architecture adheres to the standard ViT framework and incorporates 8 layers of transformer blocks for comprehensive feature extraction. To further enhance its capacity, two fully connected (FC) layers are added, leading to significant performance improvements in predicting facial beauty scores.The core network comprises 8 transformer layers, enabling the model to capture long-range dependencies within the image.The addition of two fully connected layers increases the model's capacity for more accurate predictions. Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson Correlation (PC) are utilized to measure the model's accuracy and consistency.Five-fold cross-validation is conducted, with an 80%-20% split for training and testing data in each fold.

To evaluate the effectiveness of ViT-FBP, we compared its performance against various existing methods, including both geometric feature-based and deep learning-based approaches such as LR, GR, SVR, AlexNet, ResNet-18, and ResNeXt-50. The results of this comparison are summarized in Table 4.3. ViT-FBP achieves state-of-the-art performance in facial beauty prediction, with the lowest MAE (0.1691) and RMSE (0.2149) and the highest PC (0.9534) compared to existing

Table 4.3: Performance comparison of the five-fold cross validation on the SCUT-FBP5500 dataset.

Methods	MAE ↓	RMSE ↓	PC ↑
AlexNet [80]	0.2651	0.3481	0.8634
ResNet-18 [81]	0.2419	0.3166	0.8900
ResNeXt-50 [81]	0.2291	0.3017	0.8997
CNN – SCA [71]	0.2287	0.3014	0.9003
PI-CNN [78]	0.2267	0.3016	0.8978
CNN + LDL [63]	0.2201	0.294	0.9031
ResNet-18 based AaNet [166]	0.2236	0.2954	0.9055
R3CNN [76]	0.2120	0.2800	0.9142
CNN-ER [63]	0.2009	0.2650	0.9250
GPNet [56]	0.1706	0.2225	0.9415
ViT-FBP Ours	**0.1691**	**0.2149**	**0.9534**

methods. The incorporation of Vision Transformer architecture with additional FC layers significantly enhances the model's ability to generalize across diverse data samples. Outperforms conventional CNN-based models such as AlexNet and ResNet variants by a notable margin. Demonstrates superior accuracy compared to recent approaches like GPNet, which was previously one of the top-performing methods. For this evaluation, 60% of the dataset was used for training, while the remaining 40% was reserved for testing. The testing instances were randomly selected to ensure a fair and unbiased evaluation of the model's performance. The results of the proposed ViT-FBP model were compared against various baseline methods, including both traditional machine learning and deep learning approaches. Table 4.4 summarizes the findings. The proposed ViT-FBP model achieves the lowest MAE (0.1854) and RMSE (0.2347) while attaining the highest PC (0.9519), outperforming all other methods. This demonstrates its ability to predict facial beauty with remarkable accuracy and consistency. Traditional regression-based methods such as LR, GR, and SVR show significantly higher errors, emphasizing the limitations of non-deep learning approaches in modeling complex visual features. While deep learning-based methods such as ResNet-18, ResNeXt-50, and CNN-SCA deliver competitive results, ViT-FBP surpasses them by effectively leveraging the transformer architecture for feature extraction and representation.

The Vision Transformer's ability to capture global dependencies across image patches is a key factor contributing to the superior performance of ViT-FBP compared to convolutional models like ResNet and AlexNet. The proposed ViT-FBP model establishes itself as the most effective method for facial beauty prediction in this evaluation. Its robust performance across all metrics underlines the potential of transformer-based architectures for tasks in facial image analysis.

Table 4.4: Performance comparison of different methods using the 60-40% splitting of the SCUT-FBP5500 dataset.

Methods	MAE ↓	RMSE ↓	PC ↑
LR [71]	0.4289	0.5531	0.5948
GR [71]	0.3914	0.5085	0.6738
SVR [71]	0.3898	0.5152	0.6668
AlexNet [80]	0.2938	0.3819	0.8298
ResNet-18 [81]	0.2818	0.3703	0.8513
ResNeXt-50 [81]	0.2518	0.3325	0.8777
CNN – SCA [71]	0.2517	0.332	0.878
CNN-ER [63]	0.2032	0.2683	0.9207
ViT-FBP Ours	**0.1854**	**0.2347**	**0.9519**

4.5.5 The proposed approach SPT-LSA-ViT-FBP

This section presents the results of the SPT-LSA-ViT-FBP model, which utilizes Vision Transformers (ViTs) with small-size input patches for facial beauty prediction on the SCUT-FBP5500 dataset. This architecture leverages Shifted Patch Tokenization (SPT) and Local Self-Attention (LSA) mechanisms to enhance feature extraction and improve prediction accuracy.

To evaluate its performance, we conducted a five-fold cross-validation using the SCUT-FBP5500 dataset. For each fold, the data was split into 80% for training and 20% for testing. The performance of the proposed model was compared with various baseline techniques, including both geometric feature-based and deep learning-based methods. Metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson Correlation (PC) were employed to measure accuracy and consistency. The results of this comparison are presented in Table 4.5.

The proposed SPT-LSA-ViT-FBP model achieves state-of-the-art performance, with an MAE of 0.1718, an RMSE of 0.2166, and a PC of 0.9558. This demonstrates the superior ability of the Vision Transformer-based architecture in capturing global and local dependencies for predicting facial beauty. Traditional convolutional methods like AlexNet, ResNet-18, and ResNeXt-50 deliver competitive results but are surpassed by the transformer-based approach. Advanced CNN-based methods such as GPNet show strong performance but are outperformed by the SPT-LSA-ViT-FBP, particularly in PC, reflecting better consistency and alignment with human judgments.

For this experiment, 60% of the dataset was used for training, while 40% was reserved for testing. Instances were randomly chosen for both subsets to ensure unbiased evaluation. This setup aligns with common practices in machine learning

Table 4.5: Performance comparison of the five-fold cross validation on the SCUT-FBP5500 dataset.

Methods	MAE ↓	RMSE ↓	PC ↑
AlexNet [80]	0.2651	0.3481	0.8634
ResNet-18 [81]	0.2419	0.3166	0.8900
ResNeXt-50 [81]	0.2291	0.3017	0.8997
CNN – SCA [71]	0.2287	0.3014	0.9003
PI-CNN [78]	0.2267	0.3016	0.8978
CNN + LDL [63]	0.2201	0.294	0.9031
ResNet-18 based AaNet [166]	0.2236	0.2954	0.9055
R3CNN [76]	0.2120	0.2800	0.9142
CNN-ER [63]	0.2009	0.2650	0.9250
GPNet [56]	0.1706	0.2225	0.9415
SPT-LSA-ViT-FBP Ours	**0.1718**	**0.2166**	**0.9558**

to validate the generalization capability of models. The comparison of different methods under this configuration is presented in Table 4.6, and the corresponding graphical representation is shown in Figure 4.10. Our study highlights the predominant use of supervised pre-trained models over semi-supervised or scratch-built models. This preference stems from several benefits: Reduced Training Time: Pre-trained models require fine-tuning of only a few parameters, significantly speeding up training. Simplified Task Adaptation: Modifications are typically confined to the output layer, making them versatile for various tasks. Enhanced Performance: Despite the constraints on performance improvements due to parameter limits, pre-trained models often deliver superior results.The SPT-LSA-ViT-FBP model demonstrates excellent performance with an MAE of 0.2050, RMSE of 0.2564, and a PC of 0.9488. While slightly lagging behind CNN-ER in MAE, it surpasses in PC, indicating greater alignment with human perception. The SPT-LSA-ViT-FBP leverages transformer-based architecture with Shifted Patch Tokenization (STP) and Local Self-Attention (LSA), outperforming CNN-based models on correlation metrics (PC).

The evaluation demonstrates that pre-trained models, especially transformer-based architectures like SPT-LSA-ViT-FBP, offer significant advantages in terms of accuracy, consistency, and overall performance. This highlights the transformative impact of Vision Transformers in advancing facial beauty prediction tasks.

4.6 Method Comparison

In this section, we provide an evaluation of the performance of various facial beauty prediction techniques, including proposed methods, geometric feature-based

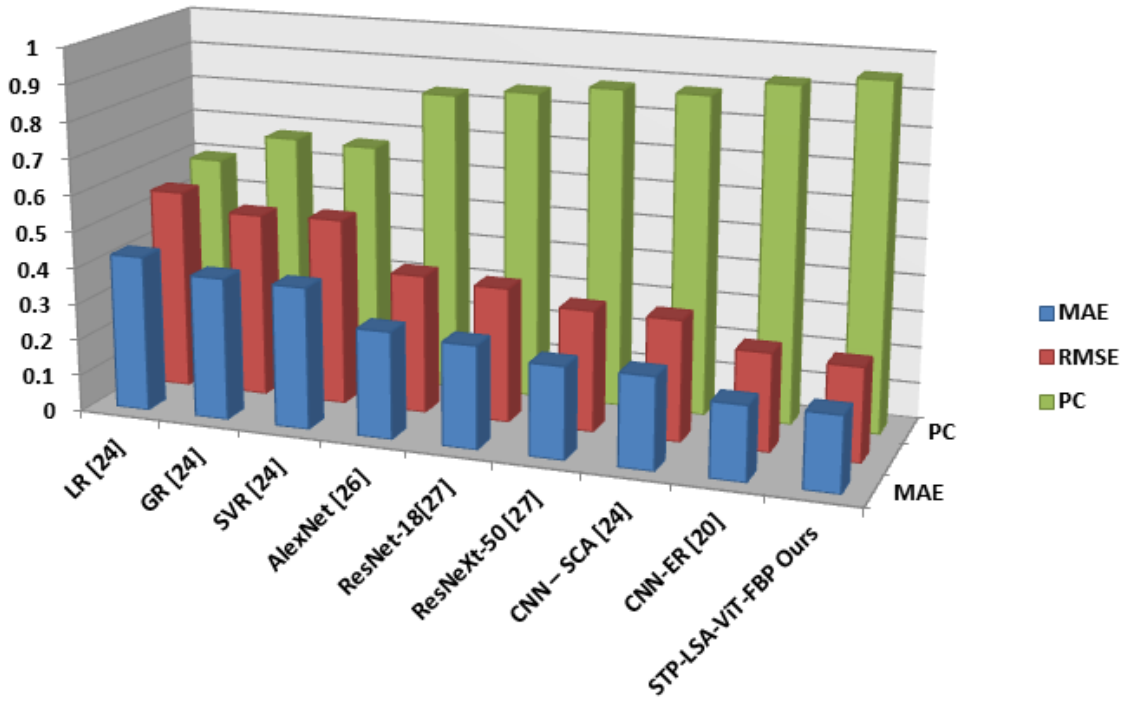


Figure 4.10: The performance comparison of different methods using the 60-40% splitting, A: Mean Absolute Error (MAE), B: Root Mean Squared Error (RMSE) and C: Pearson Correlation (PC).

Table 4.6: Performance comparison of different methods using the 60-40% splitting of the SCUT-FBP5500 dataset.

Methods	MAE ↓	RMSE ↓	PC ↑
LR [71]	0.4289	0.5531	0.5948
GR [71]	0.3914	0.5085	0.6738
SVR [71]	0.3898	0.5152	0.6668
AlexNet [80]	0.2938	0.3819	0.8298
ResNet-18 [81]	0.2818	0.3703	0.8513
ResNeXt-50 [81]	0.2518	0.3325	0.8777
CNN – SCA [71]	0.2517	0.332	0.878
CNN-ER [63]	0.2032	0.2683	0.9207
SPT-LSA-ViT-FBP Ours	**0.2050**	**0.2564**	**0.9488**

approaches, and deep learning-based approaches. This can quantify the effectiveness of techniques such as LR, GR, SVR, AlexNet, ResNet-18, and ResNeXt-50. Three main metrics were used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson Correlation (PC). Since the SCUT-FBP5500 dataset is widely used in the field of automatic FBP, we present the evaluations based on it. The creators of the SCUTFBP5500 dataset divided it into five equal sections. 80% of each section was used for training, while the remaining 20% was allocated for testing. This approach was implemented to ensure the accuracy and generalizability of the results. This makes it possible to carry out five-fold cross-validation, guaranteeing the accuracy and applicability of the findings. Figure 4.5 shows the performance comparison of the five-fold cross-validation.

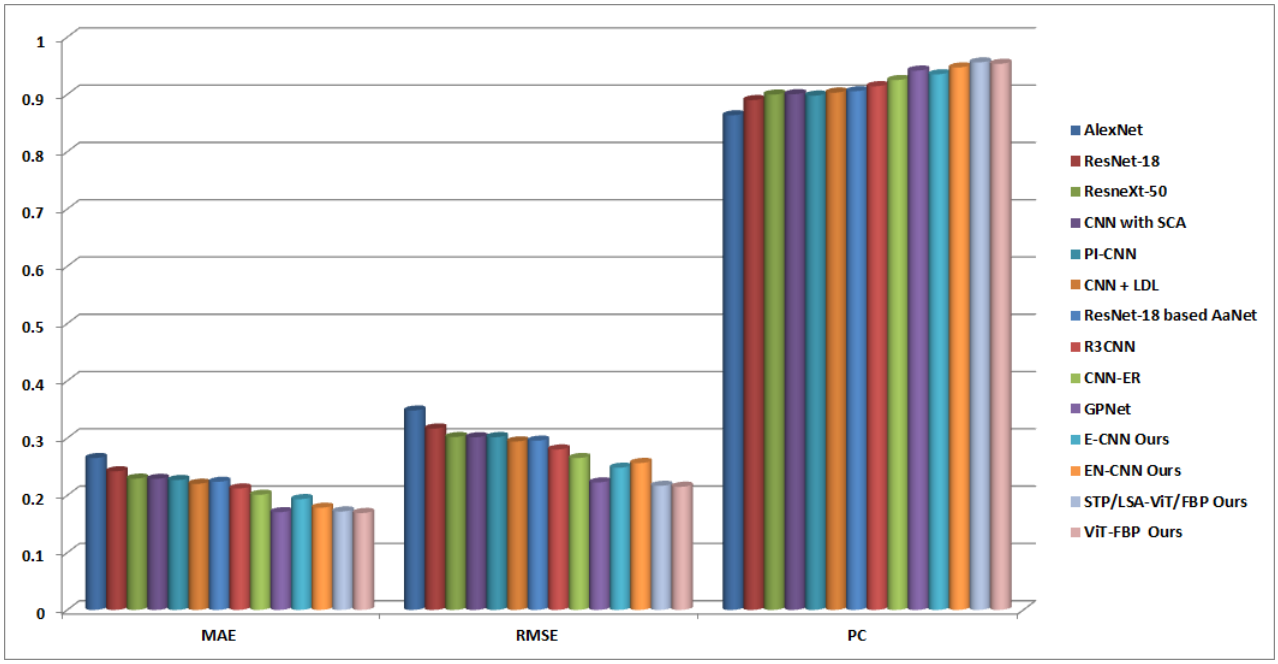


Figure 4.11: Performance comparison of the five-fold cross validation

Furthermore, the same dataset was divided into 40% for testing and 60% for training, with cases chosen at random for each group. The outcomes of this data split are shown in Figure 4.6.

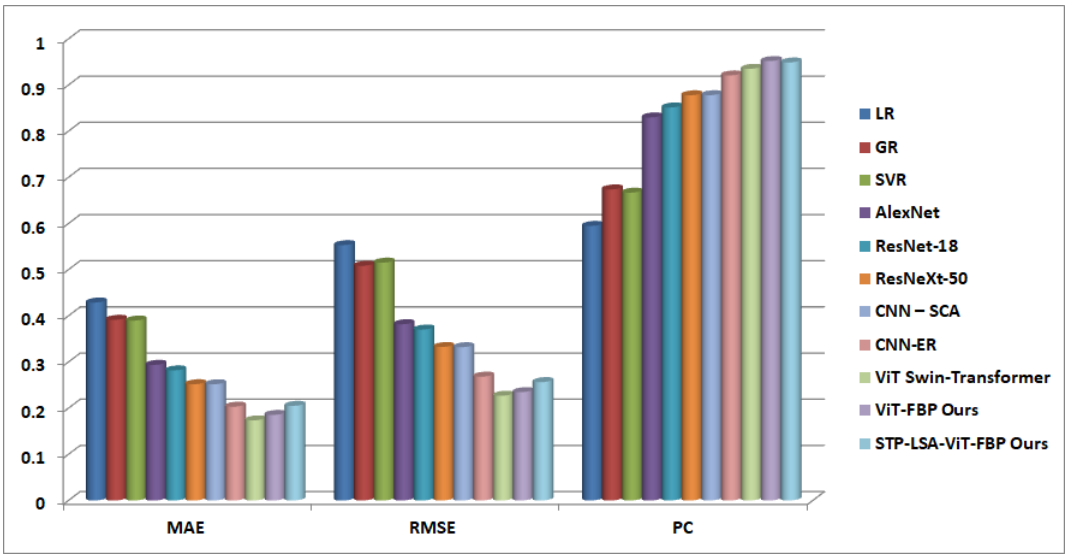


Figure 4.12: Performance comparison by 60–40% splitting of different methods

Our study's findings imply that most researchers prefer pre-trained, supervised models over semi-supervised models or models built from scratch. This preference may be attributed to a number of factors, one of which is the ease with which pre-trained models can be adjusted and the minimal number of parameter changes that are necessary. Furthermore, by modifying the output layer to meet specific job requirements, the performance of pre-trained models can be improved. Nevertheless, the number of factors often limits the performance improvement. We suggested algorithms to assist in developing more precise and successful methods for that purpose.

Our research has shown that ensemble deep convolutional neural network (DCNN) architectures outperform individual models, such as CNN-SCA, R3CNN, AlexNet, ResNet-18, and ResNeXt-50. Some of the most recent research models, such as CNN-ER, utilize a hybrid strategy for supervised learning that combines a dynamic loss function with two branches: Inception-v3 and ResNeXt-50. Using three separate DCNN pre-trained models, such as VGG16, AlexNet, and basic CNNs, is an additional strategy. PGNet utilizes a hybrid network comprising a global Swin Transformer structure and a local CNN to regulate geometric features through the regression of facial landmarks.

Our research has explored the use of deep learning techniques, such as vision transformers, to predict facial attractiveness. These methods involve analyzing facial features such as symmetry, skin texture, color, and shape to assess beauty. These approaches explore the correlation between facial attractiveness and various attributes, as well as the methodologies used for prediction. Our research has shown that vision transformer architectures yield better results than those proposed in the literature and outperform individual models, such as CNN-SCA, R3CNN,

AlexNet, ResNet-18, and ResNeXt-50. Our research on facial beauty prediction emphasizes the use of advanced technologies like deep learning, the significance of facial attributes in determining attractiveness, and the exploration of traditional concepts like the golden ratio in the context of facial aesthetics.

When compared to previous research, which usually includes gender recognition and race classification as additional tasks, these techniques perform better. Our approaches and models are able to learn complex non-linear functions and patterns in facial features, leading to more accurate predictions of facial beauty. By utilizing the SCUT-FBP5500 benchmark dataset and comparing various deep learning models, such as AlexNet, ResNet-18, and ResNeXt-50, we have demonstrated the superior performance of an ensemble-based method in facial beauty prediction. The use of our approaches in facial beauty prediction has revolutionized the field by providing more accurate and objective assessments of facial attractiveness. This advancement in technology has not only improved the accuracy of facial beauty prediction but also opened up new possibilities for industries that rely on visual appeal, such as cosmetics, fashion, and entertainment.

Chapter 5

Conclusion

5.1 limitations

Facial beauty prediction using deep learning has gained significant attention in recent years due to the increasing demand for understanding and assessing facial attractiveness. The limitations of our study on predicting facial beauty included small-scale databases. Most existing studies in this field rely on small-scale facial beauty databases, which makes it difficult to effectively model structural information for facial beauty prediction. Biased Training Data: Our models are trained on datasets that may be biased and reflect societal prejudices towards certain beauty standards. This can lead to biased predictions that reinforce societal beauty standards and exclude or discriminate against individuals who do not fit those standards. The subjectivity of beauty perception and the complexity of determining attractiveness variables remain poorly understood issues in facial attractiveness research.

These limitations highlight the need for further research to address the challenges in facial beauty prediction, such as developing more robust evaluation criteria, utilizing larger and more diverse databases, and exploring methods that reduce the reliance on subjective assessments and complex optimization procedures.
Some examples of facial beauty prediction methods that have been limited by their accuracy include:

- Ensemble of Deep Convolutional Neural Networks: While this method has shown promise in facial beauty prediction, it still faces challenges in achieving high accuracy due to the subjectivity and complexity of beauty perception.
- Dynamic Robust Losses and Ensemble Regression: This approach has demonstrated high accuracy in estimating facial beauty, but there is still room for improvement in terms of generalization performance and robustness.

- Data-Driven Facial Beauty Analysis: This method aims to improve the understanding and prediction of attractiveness in faces by focusing on prediction, retrieval, and manipulation techniques. However, the lack of consensus on relevant features and the complexity of beauty perception continue to limit its accuracy.

- Uncertainty-Oriented Order Learning: This approach targets improving the generalization performance of face beauty score prediction methods. Despite its insights into the poor generalization performance of existing methods, it still faces challenges in achieving high accuracy.

- Transfer Learning and Broad Learning System Fusion: This method has shown promise in facial beauty prediction, but it still faces limitations in achieving high accuracy due to the subjectivity and complexity of beauty perception.

These examples highlight the ongoing challenges in facial beauty prediction research and the need for further development of more accurate and reliable methods for assessing attractiveness in faces.

5.2 Future Research

Facial beauty prediction methods have gained increasing attention in the fields of computer vision, artificial intelligence, and psychology. While current approaches have made significant progress, there are several promising directions for future research to enhance the accuracy, robustness, and ethical considerations of facial beauty prediction methods. Here are some potential research directions:

- Multimodal Approaches: Combine facial features with other modalities such as voice, body language, and even biometric data for a more holistic understanding of beauty.

- Deep Learning Architectures: Explore novel deep learning architectures, including more advanced convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms, to capture intricate facial details and relationships.

- Transfer Learning and Fine-Tuning: Investigate the effectiveness of transfer learning and fine-tuning strategies to adapt pre-trained models on large datasets to specific cultural or demographic groups, ensuring the generalization of beauty standards.

Continued collaboration between experts in computer vision, psychology, and ethics will be crucial for advancing facial beauty prediction methods responsibly and ethically. Additionally, interdisciplinary research that considers social and cultural factors will contribute to the development of more inclusive and globally aware models. Future directions should focus on:

- Mitigating bias in data and models.
- Enhancing generalizability through diverse datasets and adaptation techniques.
- Improving explainability and transparency of models.
- Establishing ethical guidelines for responsible development and application.
- Exploring applications that contribute to positive societal outcomes.

5.3 Conclusion

The science behind facial beauty metrics involves the analysis of various facial features and their relationship to perceived attractiveness. Researchers have found that certain facial features, such as symmetry, averageness, and facial proportions, play a significant role in determining facial attractiveness. One of the challenges in standardizing facial beauty prediction is the lack of a universal definition of beauty. This leads to variations in beauty standards across different cultures and individuals. Furthermore, facial beauty is subjective and can vary based on personal preferences. This makes it difficult to develop a one-size-fits-all facial beauty prediction algorithm that accurately captures the perception of beauty for all individuals. This is where the use of diverse datasets, such as the SCUT-FBP5500 dataset, becomes critical. By utilizing computer vision algorithms and machine learning techniques, we can extract these facial features from images and quantify their impact on beauty ratings.
In this thesis, we prposed four algorithms based an ensemble of deep CNNs and based vision transfomers for the facial beauty prediction. These proposed was developed to predicting scores in facial beauty. The experimental findings show that our network can perform better than previous CNN baselines approaches. Experimental results showed that the proposed network achieved better performance as compared to several works available on the open literature (AlexNet, ResNet-18, ResNeXt-50, CNN – SCA, R3CNN, Semi-supervised and Vahdati at al.). These advancements have allowed for the creation of more robust and reliable prediction models, which can be applied across various industries to enhance decision-making

processes related to facial attractiveness. These advancements in deep learning and computer vision have also addressed the limitations of traditional methods in generalization and interpretability. In conclusion, the integration of deep learning and computer vision in facial beauty prediction has revolutionized the field by providing accurate and objective assessments of facial attractiveness, it improves the assessment's similarity with human judgment.

Appendices

Appendix A

Python codes

The present scripts EN-CNN are available to the public at:
https://github.com/DjameleddineBoukhari/ENCNN

The following python Codes Simple CNN Approach
The summary of Simple CNN Approach

Layer (type)	Output Shape	Param
conv2d (Conv2D)	(None, 248, 248, 32)	320
conv2d_1 (Conv2D)	(None, 246, 246, 32)	9248
conv2d_2 (Conv2D)	(None, 242, 242, 32)	25632
max_ pooling2d (MaxPooling2D)	(None, 81, 81, 32)	0
batch_ normalization (Batch Normalization)	(None, 81, 81, 32)	128
dropout (Dropout)	(None, 81, 81, 32)	0
conv2d_3 (Conv2D)	(None, 81, 81, 64)	18496
conv2d_4 (Conv2D)	(None, 81, 81, 64)	102464
conv2d_5 (Conv2D)	(None, 81, 81, 64)	102464
max_ pooling2d_1 (MaxPoo 2D)	(None, 27, 27, 64)	0
batch_ normalization_1 (Batch Normalization)	(None, 27, 27, 64)	256
dropout_1 (Dropout)	(None, 27, 27, 64)	0
conv2d_6 (Conv2D)	(None, 27, 27, 64)	102464
conv2d_7 (Conv2D)	(None, 27, 27, 64)	102464
conv2d_8 (Conv2D)	(None, 27, 27, 64)	200768
max_ pooling2d_2 (MaxPooling 2D)	(None, 9, 9, 64)	0
batch_ normalization_2 (Batch Normalization)	(None, 9, 9, 64)	256
dropout_2 (Dropout)	(None, 9, 9, 64)	0
conv2d_9 (Conv2D)	(None, 9, 9, 64)	102464

conv2d_10 (Conv2D)	(None, 9, 9, 64)	200768
conv2d_11 (Conv2D)	(None, 9, 9, 64)	200768
max_ pooling2d_3 (MaxPooling 2D)	(None, 3, 3, 64)	0
batch_ normalization_3 (Batch Normalization)	(None, 3, 3, 64)	256
dropout_3 (Dropout)	(None, 3, 3, 64)	0
flatten (Flatten)	(None, 576)	0
dense (Dense)	(None, 96)	55392
batch_ normalization_4 (Batch Normalization)	(None, 96)	384
dropout_4 (Dropout)	(None, 96)	0
dense_1 (Dense)	(None, 1)	97
Total params: 1,225,089		
Trainable params: 1,224,449		
Non-trainable params: 640		

Appendix B

Dataset

The dataset SCUT-FBP5500 analyzed during the current study is available in the github repository, https://github.com/HCIILAB/SCUT-FBP5500-Database-Release

References

[1] D. Zhang, F. Chen, and Y. Xu, Computer Models for Facial Beauty Analysis, Switzerland: Springer International Publishing, 2016.

[2] Fan, Jintu, et al. "Prediction of facial attractiveness from facial proportions." Pattern Recognition 45.6 (2012): 2326-2334.

[3] H. Knight and O. Keith, "Ranking facial attractiveness," The European Journal of Orthodontics, vol. 27, no. 4 pp. 340-348, 2005.

[4] Bruce, Vicki, and Andy Young. "Understanding face recognition." British journal of psychology 77.3 (1986): 305-327.

[5] Leyvand, Tommer, et al. "Data-driven enhancement of facial attractiveness." ACM SIGGRAPH 2008 papers. 2008. 1-9.

[6] Jain, Anil K., and Stan Z. Li. Handbook of face recognition. Vol. 1. New York: springer, 2011.

[7] D. E. Boukhari, A. Chemsa, R. Ajgou, et al., An Ensemble of Deep Convolutional Neural Networks Models for Facial Beauty Prediction, Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 27 no. 5. 2023.

[8] Tuan, Ho Nguyen Anh, Nguyen Dao Xuan Hai, and Nguyen Truong Thinh. "The Improved Faster R-CNN for Detecting Small Facial Landmarks on Vietnamese Human Face Based on Clinical Diagnosis." Journal of Image and Graphics 10.2 (2022): 76-81.

[9] J. Saeed and A. M. Abdulazeez. Facial beauty prediction and analysis based on deep convolutional neural network: a review. Journal of Soft Computing and Data Mining, vol. 2, no. 1, pp. 1-12, 2021.

[10] Yan, Haibin. "Cost-sensitive ordinal regression for fully automatic facial beauty assessment." Neurocomputing 129 (2014): 334-342.

[11] Little, Anthony C., et al. "Self-perceived attractiveness influences human female preferences for sexual dimorphism and symmetry in male faces." Proceedings of the Royal Society of London. Series B: Biological Sciences 268.1462 (2001): 39-44.

[12] Kagian, Amit, et al. "A machine learning predictor of facial attractiveness revealing human-like psychophysical biases." Vision research 48.2 (2008): 235-243.

[13] Luthfi, Muhammad, et al. "Mobile Device Facial Beauty Prediction using Convolutional Neural Network as Makeup Reference." 2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM). IEEE, 2022.

[14] Vakhshiteh, Fatemeh, Farshad Almasganj, and Ahmad Nickabadi. "Lip-reading via deep neural networks using hybrid visual features." Image Analysis and Stereology 37.2 (2018): 159-171.

[15] Nagpal, Chaitanya, and Shiv Ram Dubey. "A performance evaluation of convolutional neural networks for face anti spoofing." 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019.

[16] Xu, Jie, et al. "A new humanlike facial attractiveness predictor with cascaded fine-tuning deep learning model." arXiv preprint arXiv:1511.02465 (2015).

[17] Gan, Junying, et al. "Deep self-taught learning for facial beauty prediction." Neurocomputing 144 (2014): 295-303.

[18] Wu, Xiang, et al. "A light CNN for deep face representation with noisy labels." IEEE Transactions on Information Forensics and Security 13.11 (2018): 2884-2896.

[19] A Dosovitskiy, L Beyer, A Kolesnikov et al., "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv: 2010.11929, 2020.

[20] S. Khan, M. Naseer, M. Hayat, et al., "Transformers in vision: A survey." ACM computing surveys Vol. 54, no. 200, pp. 1-41, 2022.

[21] K. Islam, "Recent advances in vision transformer: A survey and outlook of recent work." arXiv preprint arXiv: 2203.01536, 2022.

[22] Laurentini, Aldo, and Andrea Bottino. "Computer analysis of face beauty: A survey." Computer Vision and Image Understanding 125 (2014): 184-199.

[23] Bruce, Vicki, and Andy Young. "Understanding face recognition." British journal of psychology 77.3 (1986): 305-327.

[24] El Sayed, Abdul Rahman, et al. "3D face detection based on salient features extraction and skin colour detection using data mining." The Imaging Science Journal 65.7 (2017): 393-408.

[25] Luthfi, Muhammad, et al. "Mobile Device Facial Beauty Prediction using Convolutional Neural Network as Makeup Reference." 2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM). IEEE, 2022.

[26] Kagian, Amit, et al. "A machine learning predictor of facial attractiveness revealing human-like psychophysical biases." Vision research 48.2 (2008): 235-243.

[27] Bottino, Andrea, and Aldo Laurentini. "The analysis of facial beauty: an emerging area of research in pattern analysis." Image Analysis and Recognition: 7th International Conference, ICIAR 2010, Póvoa de Varzim, Portugal, June 21-23, 2010. Proceedings, Part I 7. Springer Berlin Heidelberg, 2010.

[28] Johnston, Victor S., and Melissa Franklin. "Is beauty in the eye of the beholder?." Ethology and Sociobiology 14.3 (1993): 183-199.

[29] Cunningham, Michael R., et al. "Their ideas of beauty are, on the whole, the same as ours: Consistency and variability in the cross-cultural perception of female physical attractiveness." Journal of personality and social psychology 68.2 (1995): 261.

[30] Gaiger, Jason. "The aesthetics of Kant and Hegel." A companion to art theory (2002): 127-138.

[31] Dickerson, Adam B. "IMMANUEL KANT 1724–1804." Fifty major thinkers on education. Routledge, 2002. 60-64.

[32] Burnham, Douglas. An introduction to Kant's critique of judgement. Edinburgh University Press, 2019.

[33] Wicks, Robert. Routledge philosophy guidebook to Kant on judgment. Routledge, 2007.

[34] Cooper, David E. "Immanuel Kant (1724–1804): German Philosopher." Key Writers on Art: From Antiquity to the Nineteenth Century. Routledge, 2005. 121-125.

[35] Luo, Wei. "Aching for the altered body: Beauty economy and Chinese women's consumption of cosmetic surgery." Women's Studies International Forum. Vol. 38. Pergamon, 2013.

[36] Marlowe, Cynthia M., Sandra L. Schneider, and Carnot E. Nelson. "Gender and attractiveness biases in hiring decisions: Are more experienced managers less biased?." Journal of applied psychology 81.1 (1996): 11.

[37] Vegter, Florine, and J. Joris Hage. "Clinical anthropometry and canons of the face in historical perspective." Plastic and reconstructive surgery 106.5 (2000): 1090-1096.

[38] Bashour, Mounir. "History and current concepts in the analysis of facial attractiveness." Plastic and reconstructive surgery 118.3 (2006): 741-756.

[39] Farkas, Leslie G., et al. "Vertical and horizontal proportions of the face in young adult North American Caucasians: revision of neoclassical canons." Plastic and Reconstructive Surgery 75.3 (1985): 328-338.

[40] Farkas, L. G., J. C. Kolar, and I. R. Munro. "Geography of the nose in an attractive face: a morphometric study." Madrid: International Society of Aesthetic Plastic Surgery (1985): 36.

[41] Farkas, Leslie G., et al. "Anthropometric proportions in the upper lip-lower lip-chin area of the lower face in young white adults." American journal of orthodontics 86.1 (1984): 52-60.

[42] Farkas, L. G., C. R. Forrest, and L. Litsas. "Revision of neoclassical facial canons in young adult Afro-Americans." Aesthetic Plastic Surgery 24 (2000): 179-184.

[43] Thidar, Aye Mya, et al. "Assessing facial beauty of Sabah ethnic groups using Farkas principles." Health Sciences Research 3.1 (2016): 1-9.

[44] Atiyeh, B. S., and S. N. Hayek. "Numeric expression of aesthetics and beauty." Aesthetic plastic surgery 32 (2008): 209-216.

[45] Edler, R. J. "Background considerations to facial aesthetics." Journal of orthodontics (2014).

[46] Dunlap, Richard A. The golden ratio and Fibonacci numbers. World Scientific, 1997.

[47] Burkert, Walter. Lore and science in ancient Pythagoreanism. Harvard University Press, 1972.

[48] Meisner, Gary B. The golden ratio: The divine beauty of mathematics. Race Point Publishing, 2018.

[49] Kepler, Johannes. The harmony of the world. Vol. 209. American Philosophical Society, 1997.

[50] Huntley, Herbert Edwin. The divine proportion. Courier Corporation, 2012.

[51] Akhtaruzzaman, Md, and Amir A. Shafie. "Geometrical substantiation of Phi, the golden ratio and the baroque of nature, architecture, design and engineering." International Journal of Arts 1.1 (2011): 1-22.

[52] Prokopakis, Emmanuel P., et al. "The golden ratio in facial symmetry." Rhinology 51.1 (2013): 18-21.

[53] Thapa, Gyan Bahadur, and Rena Thapa. "The relation of Golden Ratio, mathematics and aesthetics." Journal of the Institute of Engineering 14.1 (2018): 188-199.

[54] Saeed, Jwan Najeeb, Adnan Mohsin Abdulazeez, and Dheyaa Ahmed Ibrahim. "Automatic Facial Aesthetic Prediction Based on Deep Learning with Loss Ensembles." Applied Sciences 13.17 (2023): 9728.

[55] Gan, Junying, et al. "TransBLS: transformer combined with broad learning system for facial beauty prediction." Applied Intelligence 53.21 (2023): 26110-26125.

[56] Peng, Tianhao, et al. "Geometric prior guided hybrid deep neural network for facial beauty analysis." CAAI Transactions on Intelligence Technology (2023).

[57] D. Boukhari, A. Chemsa, R. Ajgou, and M. Bouzaher, "An Ensemble of Deep Convolutional Neural Networks Models for Facial Beauty Prediction," J. Adv. Comput. Intell. Intell. Inform., Vol.27 No.6, pp. 1209-1215, 2023.

[58] Saeed, Jwan Najeeb, Adnan Mohsin Abdulazeez, and Dheyaa Ahmed Ibrahim. "An Ensemble DCNNs-Based Regression Model for Automatic Facial Beauty Prediction and Analyzation." Traitement du Signal 40.1 (2023): 55-63.

[59] Laurinavicius, Donatas, Rytis Maskeliunas, and Robertas Damasevicius. "Improvement of Facial Beauty Prediction Using Artificial Human Faces Generated by Generative Adversarial Network." Cognitive Computation (2023): 1-18.

[60] Dornaika, F. "Multi-similarity semi-supervised manifold embedding for facial attractiveness scoring." Soft Computing (2023): 1-10.

[61] Yang CT, Wang YC, Lo LJ, Chiang WC, Kuang SK, Lin HH. Implementation of an Attention Mechanism Model for Facial Beauty Assessment Using Transfer Learning. Diagnostics. 2023 Mar 29;13(7):1291.

[62] Lebedeva, Irina, Fangli Ying, and Yi Guo. "Personalized facial beauty assessment: a meta-learning approach." The Visual Computer 39.3 (2023): 1095-1107.

[63] Bougourzi, Fares, Fadi Dornaika, and Abdelmalik Taleb-Ahmed. "Deep learning based face beauty prediction via dynamic robust losses and ensemble regression." Knowledge-Based Systems 242 (2022): 108246.

[64] Dornaika, Fadi, and Abdelmalik Moujahid. "Multi-view graph fusion for semi-supervised learning: application to image-based face beauty prediction." Algorithms 15.6 (2022): 207.

[65] Gan, Junying, et al. "Facial beauty prediction fusing transfer learning and broad learning system." Soft Computing (2022): 1-14.

[66] J Iyer, Tharun, et al. "Machine learning-based facial beauty prediction and analysis of frontal facial images using facial landmarks and traditional image descriptors." Computational Intelligence and Neuroscience 2021 (2021).

[67] Wei, Wei, et al. "Assessing facial symmetry and attractiveness using augmented reality." Pattern Analysis and Applications (2021): 1-17.

[68] Xu J. Mt-resnet: a multi-task deep network for facial attractiveness prediction. In2021 2nd International Conference on Computing and Data Science (CDS) 2021 Jan 28 (pp. 44-48). IEEE.

[69] Dornaika F, Moujahid A, Wang K, Feng X. Efficient deep discriminant embedding: Application to face beauty prediction and classification. Engineering Applications of Artificial Intelligence. 2020 Oct 1;95:103831.

[70] Dornaika F, Wang K, Arganda-Carreras I, Elorza A, Moujahid A. Toward graph-based semi-supervised face beauty prediction. Expert Systems with Applications. 2020 Mar 15;142:112990.

[71] Cao K, Choi KN, Jung H, Duan L. Deep learning for facial beauty prediction. Information. 2020 Aug 10;11(8):391.

[72] Zhai Y, Yu C, Qin C, Zhou W, Ke Q, Gan J, Labati RD, Piuri V, Scotti F. Facial beauty prediction via local feature fusion and broad learning system. IEEE Access. 2020 Oct 20;8:218444-57.

[73] Gan J, Jiang K, Tan H, He G. Facial beauty prediction based on lighted deep convolution neural network with feature extraction strengthened. Chinese Journal of Electronics. 2020 Mar;29(2):312-21.

[74] Lin, Luojun, et al. "Attribute-Aware Convolutional Neural Networks for Facial Beauty Prediction." IJCAI. 2019.

[75] Zhai, Yikui, et al. "BeautyNet: Joint multiscale CNN and transfer learning method for unconstrained facial beauty prediction." Computational intelligence and neuroscience 2019 (2019).

[76] Lin, Luojun, Lingyu Liang, and Lianwen Jin. "Regression Guided by Relative Ranking Using Convolutional Neural Network (R3CNN) for Facial Beauty Prediction." IEEE Transactions on Affective Computing 13.1 (2019): 122-134.

[77] Xu, Lu, Jinhai Xiang, and Xiaohui Yuan. "CRNet: classification and regression neural network for facial beauty prediction." Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, 39Hefei, China, September 21-22, 2018, Proceedings, Part III. Cham: Springer International Publishing, 2018.

[78] Xu, Jie, et al. "Facial attractiveness prediction using psychologically inspired convolutional neural network (PI-CNN)." 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017.

[79] Liang, Lingyu, et al. "SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction." 2018 24th International conference on pattern recognition (ICPR). IEEE, 2018.

[80] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[81] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[82] Xie, Duorui, et al. "Scut-fbp: A benchmark dataset for facial beauty perception." 2015 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2015

[83] Aarabi, Parham, et al. "The automatic measurement of facial beauty." 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236). Vol. 4. IEEE, 2001.

[84] Gunes, Hatice, Massimo Piccardi, and Tony Jan. "Comparative beauty classification for pre-surgery planning." 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583). Vol. 3. IEEE, 2004.

[85] Eisenthal, Yael, Gideon Dror, and Eytan Ruppin. "Facial attractiveness: Beauty and the machine." Neural computation 18.1 (2006): 119-142.

[86] Whitehill, Jacob, and Javier R. Movellan. "Personalized facial attractiveness prediction." 2008 8th IEEE International Conference on Automatic Face and Gesture Recognition. IEEE, 2008.

[87] Gray, Douglas, et al. "Predicting facial beauty without landmarks." Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11. Springer Berlin Heidelberg, 2010.

[88] Nguyen, Tam V., et al. "Sense beauty via face, dressing, and/or voice." Proceedings of the 20th ACM international conference on Multimedia. 2012.

[89] Zhai, Yikui, et al. "Benchmark of a large scale database for facial beauty prediction." Proceedings of the 2016 international conference on intelligent information processing. 2016.

[90] Zhai, Yikui, et al. "Asian female facial beauty prediction using deep neural networks via transfer learning and multi-channel feature fusion." IEEE Access 8 (2020): 56892-56907.

[91] Xiao, Qinjie, et al. "Beauty3DFaceNet: deep geometry and texture fusion for 3D facial attractiveness prediction." Computers and Graphics 98 (2021): 11-18.

[92] Lebedeva, Irina, Yi Guo, and Fangli Ying. "MEBeauty: a multi-ethnic facial beauty dataset in-the-wild." Neural Computing and Applications (2021): 1-15.

[93] Xu, Lu, Heng Fan, and Jinhai Xiang. "Hierarchical multi-task network for race, gender and facial attractiveness recognition." 2019 IEEE International conference on image processing (ICIP). IEEE, 2019.

[94] Tong, Song, et al. "Putative ratios of facial attractiveness in a deep neural network." Vision Research 178 (2021): 86-99.

[95] Ulrich, Luca, et al. "Perspective morphometric criteria for facial beauty and proportion assessment." Applied Sciences 10.1 (2019): 8.

[96] Liu, Shu, et al. "Advances in computational facial attractiveness methods." Multimedia Tools and Applications 75 (2016): 16633-16663.

[97] Gunes, Hatice. "A survey of perception and computation of human beauty." Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding. 2011.

[98] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[99] Aamir, Muhammad, et al. "ML-DCNNet: multi-level deep convolutional neural network for facial expression recognition and intensity estimation." Arabian Journal for Science and Engineering 45.12 (2020): 10605-10620.

[100] Tan, Mingxing, and Quoc V. Le. "Mixconv: Mixed depthwise convolutional kernels." arXiv preprint arXiv:1907.09595 (2019).

[101] Persson, Ivar. "Classification of Textures Using Convolutional Neural Networks." Master's Theses in Mathematical Sciences (2017).

[102] Noura, Houacine, and Khelifa Nadia. Classification des textures par les réseaux de neurones convolutifs. Diss. Université Mouloud Mammeri, 2018.

[103] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122 (2015).

[104] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[105] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[106] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[107] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." Thirty-first AAAI conference on artificial intelligence. 2017.

[108] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning. pmlr, 2015.

[109] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[110] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[111] Xie, Saining, et al. "Aggregated residual transformations for deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[112] Byeon, Wonmin, et al. "Scene labeling with lstm recurrent neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[113] Ghorbanzadeh, Omid, et al. "Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection." Remote Sensing 11.2 (2019): 196.

[114] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." 2017 international conference on engineering and technology (ICET). Ieee, 2017.

[115] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[116] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[117] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[118] Mikami, Ayako. "Long short-term memory recurrent neural network architectures for generating music and japanese lyrics." Computer Science Department, Boston College: Boston, MA, USA (2016).

[119] Minchev, Kliment. Vision Transformer-assisted analysis of neural Image compression and generation. PhD Diss. 2022.

[120] Khan, Asifullah, et al. "A survey of the vision transformers and its CNN-transformer based variants." arXiv preprint arXiv:2305.09880 (2023).

[121] Han, Kai, et al. "A survey on visual transformer." arXiv preprint arXiv:2012.12556 (2020).

[122] Liu, Yang, et al. "A survey of visual transformers." IEEE Transactions on Neural Networks and Learning Systems (2023).

[123] Lin, Tianyang, et al. "A survey of transformers." AI open 3 (2022): 111-132.

[124] Patwardhan, Narendra, Stefano Marrone, and Carlo Sansone. "Transformers in the real world: A survey on nlp applications." Information 14.4 (2023): 242.

[125] Wu, Bichen, et al. "Visual transformers: Token-based image representation and processing for computer vision." arXiv preprint arXiv:2006.03677 (2020).

[126] He, Kelei, et al. "Transformers in medical image analysis." Intelligent Medicine 3.1 (2023): 59-78.

[127] Zhai, Xiaohua, et al. "Scaling vision transformers." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[128] Park, Namuk, and Songkuk Kim. "How do vision transformers work?." arXiv preprint arXiv:2202.06709 (2022).

[129] Ranftl, René, Alexey Bochkovskiy, and Vladlen Koltun. "Vision transformers for dense prediction." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[130] Zhou, Daquan, et al. "Understanding the robustness in vision transformers." International Conference on Machine Learning. PMLR, 2022.

[131] Paul, Sayak, and Pin-Yu Chen. "Vision transformers are robust learners." Proceedings of the AAAI conference on Artificial Intelligence. Vol. 36. No. 2. 2022.

[132] Wu, Haiping, et al. "Cvt: Introducing convolutions to vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[133] Arnab, Anurag, et al. "Vivit: A video vision transformer." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[134] Bazi, Yakoub, et al. "Vision transformers for remote sensing image classification." Remote Sensing 13.3 (2021): 516.

[135] Torrey, Lisa, and Jude Shavlik. "Transfer learning." Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI global, 2010. 242-264.

[136] Albashish, Dheeb. "Ensemble of adapted convolutional neural networks (CNN) methods for classifying colon histopathological images." PeerJ Computer Science 8 (2022): e1031.

[137] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[138] Xia, Xiaoling, Cui Xu, and Bing Nan. "Inception-v3 for flower classification." 2017 2nd international conference on image, vision and computing (ICIVC). IEEE, 2017.

[139] Peng, Shuai, et al. "More trainable inception-ResNet for face recognition." Neurocomputing 411 (2020): 9-19.

[140] Lin, Chunmian, et al. "Transfer learning based traffic sign recognition using inception-v3 model." Periodica Polytechnica Transportation Engineering 47.3 (2019): 242-250.

[141] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[142] Gulzar, Yonis. "Fruit image classification model based on MobileNetV2 with deep transfer learning technique." Sustainability 15.3 (2023): 1906.

[143] Dong, Ke, et al. "MobileNetV2 model for image classification." 2020 2nd International Conference on Information Technology and Computer Application (ITCA). IEEE, 2020.

[144] Khushi, Hafiz Muhammad Tayyab, et al. "Improved Multiclass Brain Tumor Detection via Customized Pretrained EfficientNetB7 Model." IEEE Access (2023).

[145] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, 2019.

[146] Rami Naidji, Mohamed, and Zakaria Elberrichi. "Automatic Detection of COVID-19 from Chest X-Ray Images using EfficientNet-B7 CNN Model with Channel-wise Attention." International Journal of Computing and Digital Systems 15.1 (2024): 1443-1456.

[147] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[148] Peng, Kunyu, et al. "TransDARC: Transformer-based driver activity recognition with latent space feature calibration." 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2022.

[149] Mishra, Pankaj, et al. "VT-ADL: A vision transformer network for image anomaly detection and localization." 2021 IEEE 30th International Symposium on Industrial Electronics (ISIE). IEEE, 2021.

[150] Wang, Xiang, et al. "Oadtr: Online action detection with transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[151] Lee, Seung Hoon, Seunghyun Lee, and Byung Cheol Song. "Vision transformer for small-size datasets." arXiv preprint arXiv:2112.13492 (2021).

[152] Emmamuel, Aalfin, et al. "3D-CNN method over shifted patch tokenization for MRI-based diagnosis of Alzheimer's disease using segmented hippocampus." Journal of Multimedia Information System 9.4 (2022): 245-252.

[153] Lee, Seunghoon, Seunghyun Lee, and Byung Cheol Song. "Improving vision transformers to learn small-size dataset from scratch." IEEE Access 10 (2022): 123212-123224.

[154] Dey, Ankita, et al. "Fall event detection using vision transformer." 2022 IEEE Sensors. IEEE, 2022.

[155] Zhou, Jingkai, et al. "Elsa: Enhanced local self-attention for vision transformer." arXiv preprint arXiv:2112.12786 (2021)

[156] Ma, Yiwei, et al. "Towards local visual modeling for image captioning." Pattern Recognition 138 (2023): 109420.

[157] Elorza Deias, Anne. "Face beauty analysis via manifold based semi-supervised learning." (2017).

[158] Wang, Weijie, and Yanmin Lu. "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model." IOP conference series: materials science and engineering. Vol. 324. IOP Publishing, 2018.

[159] Kumar, Satish, et al. "Performance evaluation for tool wear prediction based on Bi-directional, Encoder–Decoder and Hybrid Long Short-Term Memory models." International Journal of Quality and Reliability Management 39.7 (2022): 1551-1576.

[160] Hodson, Timothy O. "Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not." Geoscientific Model Development Discussions 2022 (2022): 1-10.

[161] Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)." Geoscientific model development discussions 7.1 (2014): 1525-1534.

[162] Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature." Geoscientific model development 7.3 (2014): 1247-1250.

[163] Benesty, Jacob, Jingdong Chen, and Yiteng Huang. "On the importance of the Pearson correlation coefficient in noise reduction." IEEE Transactions on Audio, Speech, and Language Processing 16.4 (2008): 757-765.

[164] Bottou, Léon. "Large-scale machine learning with stochastic gradient descent." Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers. Physica-Verlag HD, 2010.

[165] Loshchilov, Ilya, and Frank Hutter. "Fixing weight decay regularization in adam." (2018).

[166] Vahdati, Elham, and Ching Y. Suen. "Facial beauty prediction using transfer and multi-task learning techniques." International conference on pattern recognition and artificial intelligence. Cham: Springer International Publishing, 2020.

[167] Shi, Shengjie, et al. "Improving facial attractiveness prediction via co-attention learning." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[168] Zhang, Pingjian, and Yuankai Liu. "NAS4FBP: Facial beauty prediction based on neural architecture search." International Conference on Artificial Neural Networks. Cham: Springer Nature Switzerland, 2022.