UNIVERSITY OF CALIFORNIA SAN DIEGO

Social Impressions of Faces: Computational Modeling and Cultural Comparisons

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Cognitive Science

by

Amanda Song

Committee in charge:

Professor Virginia de sa, Chair
Professor Garrison Cottrell, Co-Chair
Professor Steven Dow
Professor Zhuowen Tu
Professor Edward Vul

2020

The Dissertation of Amanda Song is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2020

DEDICATION

To my family.

EPIGRAPH

The biggest adventure you can take,
is to live the life of your dreams.

Oprah Winfrey

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I could not have accomplished this academic discovery journey without the help and support of many people. I am deeply grateful to my committee members, Professor Garrison Cottrell, Professor Virginia de Sa, Professor Edward Vul, Professor Steven Dow, and Professor Zhouwen Tu, all of whom have given me invaluable support in this journey.

First of all, I would like to thank Professor Garrison Cottrell for taking me under his wing and his support as the co-chair of my committee. Gary has never pushed me to be anything other than fulfilling my potential: thank you! You gave me the best gift, the space to grow, and the freedom to explore the world, both inside and outside academia. Thank you for all the days and nights when we worked together, brainstorming ideas, discussing research results, writing manuscripts, polishing presentations (and also going out for a drink at music festivals)! I will carry the positive altitude and humorous spirit inherited from you to my next chapter of life.

Thanks also to Professor Virginia de Sa, the chair of my committee. You gave me a warm welcome on day zero, during my interview visit to UCSD Cognitive Science Department. I recalled that we discussed research ideas on what makes a face attractive, which later developed and extended into this whole thesis theme. When I find myself lost, you are always the person I look upon. Thank you for the feedback and advice you have given me throughout my Ph.D. Interactions with you, and your lab has been intellectually stimulating and socially fun to me.

Much appreciation goes to Professor Edward Vul. You generously offered me help when I encountered a statistical problem in my research project. Our discussion gradually developed into the most crucial chapter in my Ph.D. research. You shepherded me in my research, making sure I am asking the right question, with the correct methodology and altitude. You patiently hammered away the rough edge of my ideas and sculpted it into rich insights. I always wish that I had met you earlier, such that I could have learned more from you, in all the positive ways I could imagine.

Thanks also to Professor Steven Dow, who push me to think about the ethical issues behind this complicated story, especially in the current social context. Thank you for sharing your

knowledge and experiences with me generously and helping me bright ideas up and delivering the right message to a broader audience and the whole society. I enjoyed every conversation we had, inside and outside campus, about research and life.

My gratitude also goes to Professor Zhuowen Tu, for giving me advice on thinking about the big picture, and on how to use machine learning properly to advance scientific discovery. Your feedback on the AI fairness and inclusiveness helped me a lot in putting my research in related perspectives. You always set the highest standard on scientific study, work, and teaching, which has been a lasting inspiration to me.

I am sincerely thankful to my collaborators and co-authors: Weifeng Hu, Devendra Pratap Yadav, Fangfang Wen, Bin Zuo, Linjie Li, Chad Atalla, Bart Tam, Asmitha Rathis, Angela Yu, Shunan Zhang, Vicente Malave, Mark Chen, Suprabha Somashekhar, Nick Lin, Piotr Winkielman (and Gary and Ed). It's your hard work and generous collaboration that makes the ideas come into realization and research insights. Among them, a special thank you to Professor Angela Yu, who gave me the excellent opportunity to study and do research at one of the world's best institution's renowned cognitive science program, and who inspired my interests in face perception in the first place.

I would also like to thank the "GURU" family, Gary's Unbelievable Research Unit: Vicente Malave, Ben Cipollini, Davis Liang, Tomoki Tsuchida, Panqu Wang, Yufei Wang, Yao Qin, Yan Shu, Qianli Ma, Martha Gahl, Henry Mao; and my endearing friends and cohort inside and outside UCSD, who shared the unforgettable path with me: Sophia Sun, Aditi Mavalankar, Geelon So, Quan Vuong, Zhi Wang, Julaiti Alafate, Khalil Mrini, Bodhisattwa Majumder, Richard Gao, Tone Xu, Darlene Guo, Shuai Zhang, Chang Guo, Da Kui, Jingru Zhao, Yan Liu, and Zhujia Shi. It is your support and friendship that helped me in an immeasurable way.

Renee Yang and Mandy Zhang, it has been a great pleasure to dance and study with you, especially during the quarantine period. We are the perfect productivity squad, and we will rock-and-roll and dance together no matter where we are. Alice Xingyu Li, thanks for showing me how cool and vibrant a Ph.D. life could be and that a girl could choose her way of living to

its fullest potential and diversity. Fay He, you are the sister I choose for myself, and it's your steady support that makes me feel secure and full of power. Without sharing happiness and frustration along the way with you, life would be so colorless. Also, I would like to say special thanks to Fanji Gu and Karl Schlagenhaul, who showed me the power and fun of communicating science to the general audience vividly, and who gave me exceptional support in pre-doctoral and post-doctoral explorations.

UCSD writing hub has been a great help to me in many ways. Edward Wang, Richard Gao, you are fantastic in giving me high-quality one-on-one writing consultations. Haley McInnis, Lindsay J Depalma, and Erica Bender, thank you for your sincere advice and the inspiring and fruitful writing retreats and workshops you organized.

I would like to also thank Giulia Hoffmann at UCSD career service center for her thoughtful and professional advice and to me in critical crossroads of my career development. You have the mojo to cheer me up and your genuine emotional support means a lot to me in the past, present, and future.

During my Ph.D., I was very fortunate to explore the world outside academia through various internships and try to apply what I have learned at school to make an impact on a broader community. I want to thank my supervisor Dalar Vartanians at Tesla, who gave me full autonomy and trust to apply my computer vision algorithm in Tesla's production line and improve quality control efficiency. My sincere gratitude also goes to Vivian Sun, my supervisor at TuSimple, who taught me in her elegant and effective way, how to integrate technical acumen with business insights to help a startup succeed and innovative solutions land on the ground. She has given me invaluable guidance to make my first smooth transition from academia into the business world.

I want to give special thanks to my supervisor at the Asian Infrastructure Investment Bank, Paul Lam. He gave me an extraordinary opportunity to work in the world's youngest and most innovative multilateral development bank, with the best people one can dream of and learn from. He shows me by example how to think critically and compassionately in other people's shoes, how to jump out of the box and find a win-win solution, how to deliver with high efficiency

and communicate effectively, how to work steadily and tirelessly towards a worth-pursuing goal that's beyond ourselves. He has broadened my horizon and profoundly enlightened my thinking patterns. It's fortunate for me to find a north star in my professional career development.

I am deeply indebted to my husband, Xiaodi Hou, who inspires me to take a computational approach to study human cognition and encourages me to pursue whatever course I choose in my life. He patiently provides wise and critical counsel as well as unconditional support and love at all times. In the midst of all the responsibility he carries for his enterprise and company, he makes room and time for me and shows me the courage and grit it takes to make the impossible possible. Last but not least, I would like to thank my family, whose endless love serves as my cornerstone, especially my mom, for loving me in her unique way. Chapter 3 (predictive model), in part, is a reprint of the material as it appears in Proceedings of the 39th Annual Conference of the Cognitive Science Society (COGSCI'17). (Amanda Song, Linjie Li, Chad Atalla, Garrison Cottrell. "Learning to see people like people: Predicting social impressions of faces"). The dissertation author was the primary investigator and author of this paper.

Chapter 4 (modifAE), in part, is a reprint of the material as it appears in Proceedings of the 41st Annual Conference of the Cognitive Science Society (COGSCI'19). (Chad Atalla, Amanda Song, Bartholomew Tam, Asmitha Rathis and Garrison Cottrell. "Modifying social dimensions of human faces with ModifAE"). The dissertation author was the co-primary investigator and co-first author of this paper.

Chapter 5 (hairGAN), in part, has been accepted for publication of the material as it will appear in Proceedings of the 42nd Annual Conference of the Cognitive Science Society (COGS'20) as a poster. (Amanda Song, Devendra Pratap Yadav, Weifeng Hu, Garrison Cottrell and Ed Vul. "To Dye or Not to Dye : The Effect of Hair Color on First Impressions"). The dissertation author was the primary investigator and co-first author of this paper.

Chapter 6 (cultural model), in part, has been accepted for publication of the material as it will appear in Proceedings of the 42nd Annual Conference of the Cognitive Science Society (COGSCI'20)., (Amanda Song, Weifeng Hu, Devendra Pratap Yadav, Fangfang Wen, Bin Zuo,

Garrison Cottrell and Ed Vul. “Do you see what I see? A cross-cultural comparison of social impressions of faces”). The dissertation author was the primary investigator and co-first author of this paper. Chapter 6, in part is also currently being prepared for submission for publication of the material. (Amanda Song, Weifeng Hu, Devendra Pratap Yadav, Fangfang Wen, Bin Zuo, Garrison Cottrell and Ed Vul.) “Mediating mechanisms of facial impressions: cultural universals and idiosyncrasies.” The manuscript’s title may be subject to change. The dissertation author was the primary investigator and author of this material.

VITA

2008-2012 B.S in Biology Science, Fudan University

2014-2017 M.S in Cognitive Science, University of California San Diego

2014-2020 Ph.D. in Cognitive Science, University of California San Diego

PUBLICATIONS

Amanda Song, Weifeng Hu, Devendra Pratap Yadav, Fangfang Wen, Bin Zuo, Garrison Cottrell and Ed Vul. "Mediating mechanisms of facial impressions: cultural universals and idiosyncrasies.' Manuscript under preparation. Title may be subject to change.

Amanda Song*, Weifeng Hu*, Devendra Pratap Yadav, Fangfang Wen, Bin Zuo, Garrison Cottrell and Ed Vul. "Do you see what I see? A cross-cultural comparison of social impressions of faces". Proceedings of the 42nd Annual Conference of the Cognitive Science Society, 2020 (* equal contribution)

Amanda Song*, Devendra Pratap Yadav*, Weifeng Hu, Fangfang Wen, Bin Zuo, Garrison Cottrell and Ed Vul. "To Dye or Not to Dye: The Effect of Hair Color on First Impressions". A poster of the 42nd Annual Conference of the Cognitive Science Society, 2020 (* equal contribution)

Amanda Song, Linjie Li, Chad Atalla and Garrison Cottrell. "Learning to see people like people: Predicting social impressions of faces". Proceedings of the 39th Annual Conference of the Cognitive Science Society, 2017 (oral presentation)

Linjie Li, Vicente Malave, **Amanda Song** and Angela J. Yu "Extracting Human Face Similarity Judgments: Pairs or Triplets?". Proceedings of the 38th Annual Conference of the Cognitive Science Society, 2016

Shunan Zhang, **Amanda Song** and Angela J. Yu "A Bayesian hierarchical model of local-global processing: visual crowding as a case-study". Proceedings of the 37th Annual Conference of the Cognitive Science Society, 2015

Amanda Song, Ai Koizumi and Hakwan Lau "A Behavioral Method to Manipulate Metacognitive Awareness Independent of Stimulus Awareness". Behavioral methods in consciousness research, 2015.

ABSTRACT OF THE DISSERTATION

Social Impressions of Faces: Computational Modeling and Cultural Comparisons

by

Amanda Song

Doctor of Philosophy in Cognitive Science

University of California San Diego, 2020

Professor Virginia de sa, Chair
Professor Garrison Cottrell, Co-Chair

At first sight of a new person, people quickly form impressions on how attractive, trustworthy, and warm the person looks. Despite the dubious accuracy of these first impressions, people rely on them to navigate social interactions regarding interpersonal relationships, political voting decisions, and financial choices. As we progress into a digital era with frequent social media usage and cultural exchanges, it is crucial to understand the universal principles and cultural idiosyncrasies of impression formation from face images.

In this thesis, we examine social impressions of faces through the lens of computational modeling, psychological experiments, and cultural comparisons. First, we develop a model that

automatically predicts human social impression judgments using neural networks. Building on the predictive model, second, we build a generative model that can change faces holistically to augment or decrease specific characteristics, such as attractiveness and aggressiveness of a face. Third, we examine how specific physical attributes, such as hair color, affect impressions of faces, using a GAN model, and psychological experiments. Finally, we conduct a large-scale cross-cultural study, using 18 traits related to approachability, youthful-attractiveness, and competence evaluation, with Caucasian and Asian participants rating Caucasian and Asian faces. We investigate the mediating factors behind impression formation and estimate how high-level facial features such as age, eyeglasses, and smiles are judged similarly and differently by people from two cultures.

Overall, our work provides a computational framework to predict and modify faces, which is practically useful in real-life scenarios regarding optimal self-image presentation. Our model lays the foundation for the quantitative study of first impressions. Our psychological and cross-cultural studies reveal the universality and culture-specific judgments mediated by high-level features that affect first impressions. These findings motivate future research in social psychology to understand the deeper cultural roots behind the differences in first impression formation and be aware of the potential bias toward certain social groups.

Chapter 1

Introduction

1.1 General background

Humans are social animals who make use of every possible cue to navigate in complex social interactions. Despite the warning to never judge a book by its cover, when people encounter a stranger for the first time, they read from a person and spontaneously form impressions about the person's attractiveness, trustworthiness, social status, etc.

Among the multiple cues people use to form impressions, such as the clothing, body language, body shape, and weight [76], voice, and most importantly, behavior, in this thesis proposal, we focus on first impressions drawn from facial photos alone for two reasons. First, human faces contain rich social signals about emotions, intentions, and attitudes, and face perception has been well-studied. Second, first impressions in the digital age are primarily reduced to our impressions of facial photos. Every second, people are uploading and browsing billions of pictures on-line. They are the most effortless food for the mind, as Lippmann has remarked in his book[108]. Profile photos on websites are often the first thing we learn about a stranger. Employers regularly check professional sites such as LinkedIn to learn about a job candidate. In dating websites and mobile apps, people will quickly learn about each other through photos.

Although the spontaneously formed impressions are not objectively true[130], these snapshot subjective evaluations of others have a profound influence in various aspects of our

lives, regarding who we want to hire, to be friends with, to date, to vote, to promote and even to sentence for legal punishments [184]. Given the profound social consequences of facial impressions, it is of great scientific value to understand how people form impressions of each other based on facial images, and to discover the shared assumptions/ inferences that perceivers have about the social evaluation of faces.

Research on first impressions has a long history [7]. Regardless of the dubious accuracy of first impressions, psychologists have found that first impressions can be formed in as short as 50 milliseconds after exposure to a face photo[200, 10] and there is a substantial agreement [28] in these judgments across people, across cultures [213, 34] and ages [165, 29]. In Chapter 2, we will review previous work in first impression studies in depth from multiple disciplines, including psychology, social science, machine learning, and economics.

1.2 Technological goals

With the progress in deep learning research, there has been tremendous improvement in various vision-related tasks. However, most of the existing computer vision work on faces focuses on modeling the objective properties of face perception, such as identity recognition, expression, pose, and landmark detection. Relatively less attention has been paid to the subjective impression of faces. The previous main work has been on attractiveness rather than other traits such as trustworthiness. In contrast, there has been little work aimed at systematic evaluation and prediction of a wide range of impression traits.

1.2.1 Predicting social perceptions carried by facial photos

To bridge this gap, in Chapter 3, we proposed state-of-the-art neural-network-based prediction models of 40 social traits such as attractiveness, trustworthiness, intelligence, friendliness, and aggressiveness. We trained the model on a comprehensive and naturalistic dataset and compared the deep-learning based representations with traditional human-engineered facial features in predictive power for various traits. This model offers human-like predictions on a large scale

and can be applied to different datasets to examine the relationship between impressions and face information, and can offer guidance for impression management purposes, such as helping people pick up the most trustworthy photo among hundreds of selfies.

1.2.2 Synthesizing and changing social impressions in faces

In Chapter 4, we move further from the prediction of impression traits to modification of them. We first apply the prediction model on a larger scale dataset, CelebA, and predict human subjective opinions of these facial photos. We then use the predicted social impression ratings on faces as our augmented dataset to supervise a new model to make modifications to the images, such as to make a photo look more trustworthy. We develop an auto-encoder based neural network that can modify face impressions in a continuous and multi-dimensional manner. With this model, we can enhance or reduce the perceived value of a face in one or more specified dimensions (including both subjective impression traits as well as relatively more objective traits such as gender).

With a system capable of predicting and changing the social perceptions elicited by facial photos, social psychologists have more power than ever to parameterize face stimuli and to visualize the desired combination of factors of interests.

Apart from holistically modifying a face to enhance its perceived impression values, we can also focus on changing specific features of the photo, such as hair color, and see how tweaking one factor influence the perceived subjective impression values. In Chapter 5, utilizing the state-of-the-art GAN(Generative Adversarial Network)s, we probe how hair color affects people's impression on attractiveness, intelligence, and trustworthiness of a face photo. This methodology can be applied to more facial features of interest, such as wearing eye-glasses, having a beard, wearing lipstick, and can allow future researchers to examine carefully how their feature of interest would affect impression formation.

1.3 Scientific goal

Universality and idiosyncrasies are two sides of the same coin for first impressions. Universality, the agreement people share on a specific impression, may reflect the influence of our shared evolutionary history on our perception; idiosyncrasies, the disagreement people bear on a particular impression, on the other hand, may reflect the impact of our unique environment, including culture, personal history, and our bias on our impression formation.

Culture shapes our social norms, expectations, and values. For instance, East Asians have been characterized as being more collective and holistic, whereas Westerners have been more individualistic and analytic [75, 133]. It should come as no surprise that the cultural background of the perceiver also shapes the impression formed from faces. The social importance of the cross-cultural study of first impressions may be increasingly immense, given the preponderance of face-to-face international interactions over video conferencing and social media these days.

1.3.1 Cultural comparison of facial impressions

In Chapter 6, we build a large-scale cross-cultural dataset of facial impressions. The original US 10K Adult Database is mainly composed of Caucasian viewers' ratings of Caucasian faces. To increase the cultural diversity and inclusiveness of the dataset in terms of both observer ethnicity and face ethnicity, we conducted a study using 18 traits related to approachability, youthful-attractiveness, and competence evaluation, with Caucasian and Asian participants rating Caucasian and Asian faces. This database is suitable for the analysis of cultural differences of social evaluations of faces and suitable for training a more inclusive and representative machine learning system.

1.3.2 Mediating factors of facial impression formation

After the large scale dataset is collected, we estimate how high-level facial features such as age, eyeglasses, smiles, etc. influence impressions for Caucasian and Asian raters.

It offers a clear and interpretable understanding of how people use various facial features to form impressions. In addition to cultural universals in how high-level facial characteristics relate to impressions, we find the following cultural differences: (a) Asians give overall lower positive ratings to faces compared to Caucasians. (b) Caucasians and Asians agree more on approachability-related traits, but (c) less on competence-related traits, as Caucasians rate older people as more successful and responsible whereas Asians do the opposite. Our findings provide insights into the mediating factors of cross-cultural perception of facial traits and high-level facial features critical to future research.

1.4 Ethical implications

The application of the first impression research is not restricted to the set of issues listed above. Native application areas include social robots, artificial agents, and human-computer interfaces. Given its wide application in real life, the study of social perception of faces has a range of ethical obligations.

In a society that is undergoing tremendous economic, cultural, and societal changes, studies of first impressions can help us reveal and understand the potential source, types, and degrees of shared social consensus and cultural specific opinions towards different social groups (gender, age, and cultural groups). Our computational model captures and explains the differences and similarities of the social perception of faces across cultures. The cross-cultural dataset we collected brings in broader diversity to the field and will benefit future researchers to train an inclusive machine learning predictive and modification algorithm for practical application such as profile photo recommendations and culturally-customized modifications. Our computational model and large-scale dataset can also be used to illustrate the potential biases, raise people's awareness of it, educate people how bias creeps into the system, rather than use the algorithm to exacerbate the problem blindly.

It's crucial to have honest conversations about bias, diversity, and inclusion and elevate

the voices of potentially discriminated groups. Being aware of the potential biases for various demographic groups is the first step to build a fair and inclusive society. In this thesis' context, it refers to the potentially biased impressions towards specific individuals or social groups. Being aware and vigilant that although our brain will automatically form impressions of people, these impressions are not necessarily accurate will help us use rationality to guide us counterbalance our spontaneous responses, and to mitigate stereotypes and systematic biases.

1.5 Disclaimer

Innovations in artificial intelligence systems have advanced our society, improved productivity, and rendered new opportunities in science and industry. But at the same time, they also bring up new challenges and emerging ethical issues. It requires the whole society to make an effort to ensure that we employ AI in a fair, accountable, transparent, and ethical way[2].

In this thesis, we conducted an interdisciplinary study of social perception of faces, developed computational models to probe the cultural influence, and perceptual basis of social impressions. It is essential not to confuse the impressions we draw from faces with other people's actual personalities. The computational models we developed based on human subjective rating data reflect more about the participants' own biases (if any), which are embedded in our society today, rather than the accurate attribution and personality of others. Machine learning models built on top of this system should be used with caution. It can serve as a tool to reveal and quantify the potential bias inside the system, which requires future endeavors to mitigate. In no circumstance should it be used to make inaccurate and improper claims about particular social groups or individuals.

We take one small step in increasing the diversity of human social psychological databases with our cross-cultural data collection. Yet the current dataset is still not exhaustive, and it is beneficial to include more samples from a broader range of populations to increase further the diversity that reflects our society.

Chapter 2

Related work

Research on first impressions has a long history. Over the past decades, researchers in different fields (face perception, social and cognitive psychology, computer science, behavioral economics, and human-computer interaction) have made substantial theoretical and empirical advances that shed light on the fundamental questions concerning the study of first impressions. As the study of social impressions is a multi-disciplinary venture, we aim to review the work from multiple fields. Our review is non-exhaustive, but we aim to cover a wide range of findings and point out the internal relationships between them.

In this chapter, we first discuss what we mean by the first impression, followed by a review of findings in social psychology demonstrating the critical social consequences of them, the mixed evidence on the accuracy of these impressions, with an emphasis on behavioral economics findings that tie first impressions to people's decision making in social interactions.

Next, we summarize the psychology and social science findings, starting from face perception research on the visual cues that subserve first impressions, and then move on to the data-driven psychological models. After reviewing the perceptual cues, we then consider the roles of non-perceptual contexts in impression formation.

We then review the cultural comparison findings on first impressions, and lastly, we outline previous work on computational models of facial attribute prediction, and image editing and modification.

2.1 Social consequence and accuracy of first impressions

Facial first impressions refer to the rapid, far-reaching social inferences from visual facial cues [174]. People share a large degree of consensus over the impressions [28, 110, 188] which take takes no more than 100 milliseconds to form [185]. Additional time will increase people's confidence but doesn't change the evaluation outcome. This study suggests that first impression formation is an automatic and spontaneous process and albeit one that fundamentally touches our daily life.

2.1.1 Consequence

Social inferences made from faces are as consequential as they are ubiquitous. Below, we will review the social consequences of first impressions in dating, political decisions, military ranking, and legal domains, with a particular emphasis on economic areas.

Although facial attractiveness is a contributing factor in mating choice, other social attributes such as assertiveness, extroversion, and intelligence have also been shown to play a role[131]. This conclusion holds even after controlling for facial attractiveness.

In the political domain, the perceived competence of faces is demonstrated to predict election results in the US[184]. Strikingly, these findings hold across cultures[103, 118] and even small children can predict election results from faces with above chance accuracies[6]. Similar studies have shown that other social dimensions of a face can also predict election outcome, such as trustworthiness[25], sociability[23], stereotypicality[129]. A recent study has found that people can even infer whether a politician is corrupt or not from their face photos [107].

In the military domain, perceived dominance of face photos of cadets predicted their later military ranking attainments [119, 126].

Unfortunately, the influence of first impressions has extended to the legal domain. Studies have found that the defendants' physical attractiveness and assigned length of sentencing are negatively correlated[105]. Other studies have shown that defendants with untrustworthy looking

faces or stereotypically criminal-looking faces are more likely to receive guilty verdicts[139]. Perceptions of untrustworthiness predicted death sentences for convicted murderers[201]. These results suggest an alarming bias in the criminal justice system.

Economic decision making

The influence of facial impressions has been shown both in strategic games in the lab and in real-world financially-relevant scenarios, with both correlation studies and causation analysis. For in-lab experiments, one frequently used paradigm is trust game[12], which was developed to approximate the everyday situation involving trust and reciprocity. In a trust game of two participants, both are given some quantity of money in the beginning of the game. The first player is told that he must send some money to the second player, although the amount could be zero. Both players are told that whatever the first player sends will be tripled by the experimenter. The first player chooses a value; then, the second player receives a tripled amount. The second player then needs to give some amount of the now-tripled money back to the first player, although the amount could be zero [12].

One natural question is how initial impressions can influence trust. Studies have found that individuals with less-trustworthy looking faces receive less trust in trust games [24, 193]. The facial trustworthiness "premium" remains significant even when a reputation signal is available[143]. The trustworthiness premium is also found in zero-sum game studies, in which results have shown that a threatening face has little influence on wagering behavior, but faces relaying trustworthiness do[160]. In another trust game study[180], researchers investigated the causal link of facial trustworthiness and behaviors by allowing players to choose a computer-generated face to represent them in the game. They have found that players choose faces that look more trustworthy, and the selected avatars have an influence on strategic choices. In a similar vein, other studies have found that positive facial expressions serve as signals of honesty in a one-shot Prisoner's dilemma [141]. In another empirical study[46], Eckel and Petrie found that a face has information value in strategic decisions in that players are willing to pay the price

to see their counterpart's picture. Several studies aim at answering the question: to what extent and under what conditions people can detect cooperativeness and other social signals from faces. We will review them in the accuracy of first impressions section 2.1.2.

In real-world scenarios, studies have found that competent and dominant looks factored into CEO compensations [60, 149]. Ratings of competence and leadership predicted the number of profits that the CEOs' companies made[149]. Baby-faceness is positively correlated with black CEOs' success, but the trend is opposite for white male CEOs [115].

2.1.2 Accuracy

Do people's first impressions of others reflect some "kernel-of-truth"? The evidence is highly mixed [184, 155].

Studies of personality impressions have shown that some personality traits can be inferred with above-chance accuracy from faces, especially of extroversion, agreeableness, and neuroticism[137]. Another large body of research has claimed that people are able to accurately infer a variety of behavioral tendencies such as sexual orientation[151, 154], religious belief[152], political orientation[150], corruptibility[107], and even criminal behavior[192]. However, many of these studies didn't control for the more obvious factors, such as the gender, age, and ethnicity of a person, which could be directly used to make better than chance guesses. Olivola and Todorov have shown that judgments based on only the factors mentioned above and base-rate information about the distribution of these factors in the population are more accurate than judgments made based on facial impressions[130].

There are a growing number of studies in behavioral economics that aim at establishing the link between facial impressions and strategic decision making. Studies have shown that humans can predict their opponents' decisions above chance in a prisoner's dilemma game after a face-to-face interaction [141, 169]. Studies have shown that males with greater facial width are more likely to exploit the trust of others, and conversely, other players are less likely to trust male counterparts with higher facial-width ratio[172]. One study found that pictures taken at the

moment of decision making is diagnostic of cooperating/defect behavior in a one-shot prisoner's dilemma, but photos taken before the actual game have no predictive power[195]. Another study found that people can assess altruism correctly based on impressions[128]. On the other hand, another line of studies has found that people are not accurate when perceiving the trustworthiness in economic games like prisoner's dilemma and the stag game [196, 47].

There is also an investigation into the accuracy of intelligence perception that showed mixed results. Zebrowitz and her colleagues have found that the relationship between perceived intelligence and actual intelligence measured by IQ tests depends on life span, with higher accuracy in childhood but not in late adulthood[209]. Also, perceived intelligence is mediated by facial attractiveness, and the accuracy only holds for targets who are below average in facial attractiveness[212].

Finally, it is worth noting that the accuracy results depend on the actual face stimuli used, the experimental paradigm, and the measurement adopted.

Even if there is a valid signal in facial impression to predict people's characteristics, it is likely to be driven by self-fulfilling prophecy, namely that people will behave in a way expected by others and the society, therefore fulfill the expectation in the end. Although people's subjective impressions may not be accurate, we nonetheless share specific consensus and rely on this consensus to navigate in social contexts. Therefore, it is essential to illustrate what the shared subjective impressions of faces are, and what the mapping from facial image to the perceptual impression space is.

2.2 Face features related to impression formation

When people consistently form the same impression from faces, one natural question that people tend to ask is: what makes a face look like trustworthy/ attractive/ aggressive, and so on? In other words, what morphological characteristics trigger specific impression trait perception?

2.2.1 Hypothesis driven approaches

A large body of studies has aimed at identifying the facial cues that lead to corresponding impressions. Back in 1954, Paul F. Secord and his colleagues conducted a series of experiments to probe the relationship between personality traits and faces. They found that people with thin lips and wrinkles at the corners of the eyes are perceived as distinguished, intelligent, and determined [161]. Among physiognomic facial features (e.g., the distance between eyes, the fullness of lips), the ratio of facial width-to-facial-height (fWHR) is positively correlated with perceived and actual aggression[22].

Next, we will review the most obvious set of cues that we immediately notice from faces, such as age, gender, ethnicity, and emotional expressions. These cues are, not surprisingly, related to first impressions.

Facial age

When people age, there are structural changes in the face shape, and there are texture changes in the skin (wrinkles, unevenness in skin color). These biomarkers are picked up by our visual system to infer the "biological age" and maturity of a person[27]. Face maturity, and apparent age play a role in impression formation[13]. Research has shown that people who are perceived as older are rated as less attractive, less likable, less distinctive, less growth-oriented, and less energetic[45]. On the opposite end, people with babyfacedness look (large eye size, thinner eyebrows, and round eyes[211]) are perceived as physically weak, naive, honest and warm[14].

Facial gender

Female and male faces are structurally different, and people can accurately infer the biological gender of faces[19, 20]. Female faces look more babyfaced[57] and bear a more considerable similarity than male faces to the happy expression[1]. Feminine-looking females are perceived as attractive, and both feminine-looking and masculine-looking males are recognized

as attractive, and masculinity is associated with dominance[132].

Facial race

Although race is not necessarily grounded based on a clear biological definition, it is a concept frequently used in social categorization, and it influences impression formation. Previous research has found that black faces are perceived as more athletic, less reserved, and more masculine; Asian faces are perceived as more reserved, less dangerous, and more competent; and white faces are perceived as in between[210] in these traits.

Face attractiveness

Among the various first impression traits, facial attractiveness has received the most attention. As early as 1879, Galton found that the average face is more attractive than the individual ones that it consists of. This beauty-in-averageness effect has been confirmed by later studies[100]. Apart from averageness, facial symmetry also contributes to facial attractiveness[145, 84]. There have been evolutionary arguments that facial attractiveness is a reliable cue for good genes, indicative of health and fertility[159]. Facial attractiveness itself plays an important role in other impression formation. The "halo effect", which is also known as the "what-is-beautiful-is-good" stereotype[40], refers to the tendency for people to rate attractive individuals more favorably for their personality traits, such as sociability, popularity, intelligence and competence[116, 44]. Attractive people are perceived as being psychologically adapted, while unattractive people are perceived as less willing to cooperate[167]. Facial typicality has been shown to contribute to facial attractiveness and trustworthiness in [167].

Facial expression

Dynamic facial expressions are also a valuable source of influence on first impressions. Smiling faces look affiliative, submissive, and trustworthy while angry faces look dominant[72]. Smiling faces also are perceived as more sociable[17]. Faces with surprise and fear emotions look less dominant[125]. Although the relationship between emotions and first impression, at

first sight, seem trivial and obvious, at a deeper level, it reveals an interesting fact that people are using transient and dynamic cues (emotional expression) to infer relatively long-lasting traits (personality and impressions). Secord and his colleagues first proposed the term "temporal extension" [161] to explain this phenomenon, and Zebrowitz and her colleagues later developed it as "over-generalization" theory[210]. Additional evidence supports the theory: expressionless faces rated as threatening were also classified as "angry" by a neural network trained on emotional and neutral faces[157]. The over-generalization argument also fits with the ecological theory of social perception, which argues that face perception has adaptive value, and recognizing an angry or aggressive face is crucial for survival[120]. Therefore, people tend to read "emotion" and "impression" from structurally similar faces, regardless of whether the face actually carries the specific emotions. The over-generalization hypothesis has been applied to gender, race, and familiarity[208].

To summarize, previous research has identified the various facial features that subserve impression formation. However, thousands of facial features could be relevant, and it is unrealistic to test them one by one. Moreover, some features are hard to verbalize or impossible to predefine.

2.2.2 Data-driven approaches

The early eigenface model

In 2005, Brahnam and his colleagues built a computational model based on principal component analysis (eigenface)[18] that can predict facial impression(classified as high or low on some trait) and synthesize faces. However, the paper did not analyze the facial configuration transformation underlying the perception of each attribute.

3D layer scanned face space model

To bridge this gap, and to propose a theoretical framework for analyzing multiple impression trait, Oosterhof and Todorov[132] collected a dataset of numerous social trait ratings and then used principal component analysis to capture the critical dimensions of first impressions.

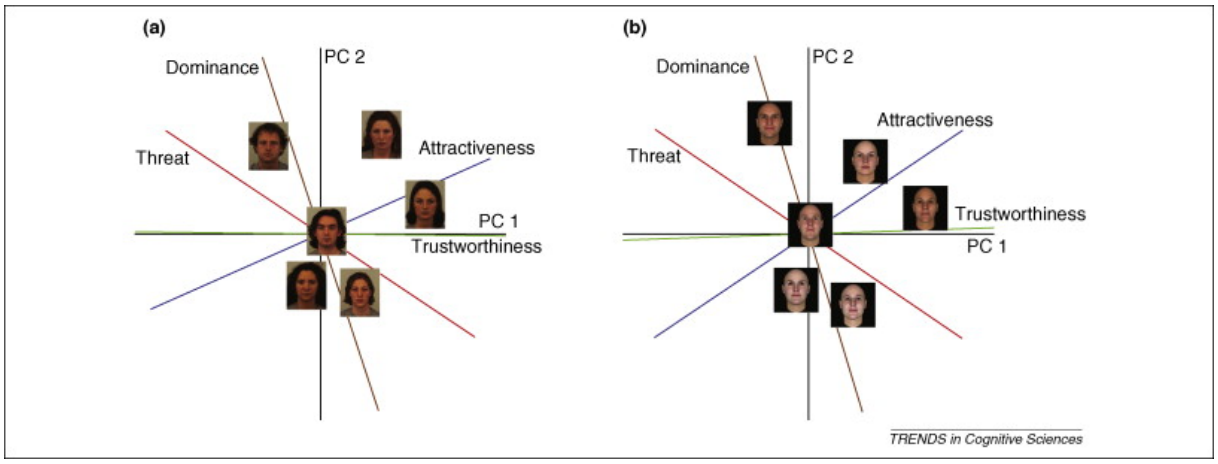


Figure 2.1. The structure of face evaluation. Principal components analysis of trait judgments. Extracted from Oosterhof and Todorov[181].

In this study, they first collected 55 participants' unconstrained description of impressions of 66 facial photos. Then they clustered these impression words based on semantic meanings and the frequency they appeared and reduced the focus to 14 traits. Through principal component analysis, they discovered two key dimensions of first impressions that capture 82% of the variance. The first key dimension is best described as trustworthiness, and the second is approximated by dominance in Fig2.1. What is worth noting is that very similar patterns emerge from both natural faces and computer-generated faces, indicating the robustness of the dimensions. The 2D structure of face evaluations aligns well with previous findings in social psychology[52].

Oosterhof and Todorov further use the statistical face model based on 3D laser scans of faces to visualize the changes in a face that affect specific impression perception and to parameterize the degree of desired impressions. They represent the shape and reflectance of the faces with PCA to extract the components that account for most of the variance in face shape. In this way, they build a 50-dimensional face space. They then used the mean ratings of specific traits to find the vectors in the 50-dimensional space that represent changes in the corresponding characteristics. For instance, they found the vector of trustworthiness and attractiveness and defined a plane with these two vectors. In the original mean ratings, the correlation between trustworthiness and dominance is positive. They then either keep the original vector, or rotate

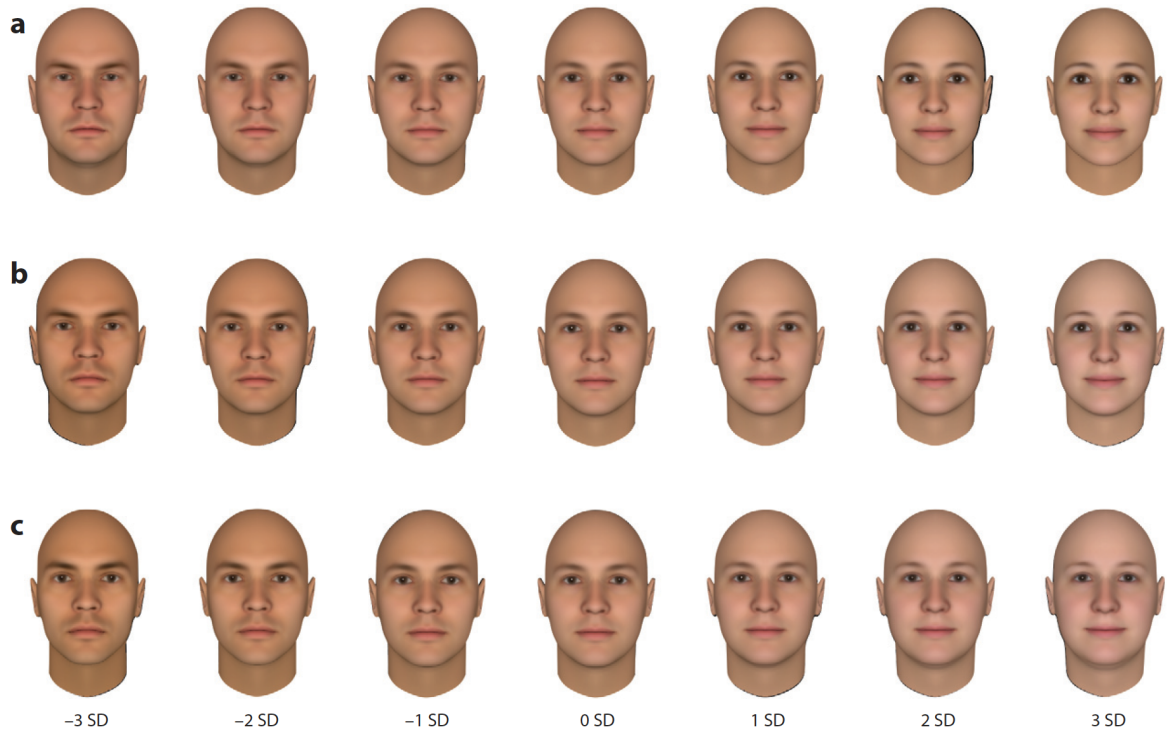


Figure 2.2. A model of trustworthiness with varied degrees of control for attractiveness. (a) Trustworthiness is positively correlated with attractiveness. (b) Trustworthiness is orthogonal to attractiveness. (c) Attractiveness have been subtracted from trustworthiness [182].

the attractiveness vector to make it orthogonal to trustworthy, or to subtract attractiveness vector from trustworthiness vector to create a negative correlation in Fig 2.2. Lastly, they validate that the faces generated by this model successfully manipulated trustworthiness and attractiveness perception [182].

Ambient image approach

A major critique of this computer-generated face space method is that the computer-simulated faces lack naturalistic variation in real faces; therefore, it has the risk of neglecting important variations that play a role in real life. To address this challenge, a group of researchers advocated the ecological value of using natural images and highly variable face photographs in face perception studies [21, 175]. Using naturalistic facial photographs, Jenkins and his colleagues have found that the within-person variability is greater than across-person variability

Figure 2.3. Morphed continua for age, sexual dimorphism, attractiveness, intelligence, confidence, trustworthiness and dominance. Extracted from Sutherland[175]

in facial attractiveness [81]. Similar results have been found for competence, creativity, cunning, extroversion, meanness, smartness, and trustworthiness judgments[186]. Using 1000 ambient facial photos collected from the Internet, Sutherland and her colleagues have found an additional dimension of youthful-attractiveness[175] besides the original two key dimensions of trustworthiness and dominance[181]. Furthermore, with averaging and morphing techniques, they visualized the transformation maps of age, sexual dimorphism, attractiveness, intelligence, confidence, trustworthiness, and dominance[175]. See Fig2.3 for details.

Reverse correlation and bubble methods

Reverse correlation is another method employed by psychologists to visualize the changes underlying social perception of faces[42], stereotypes of groups[77], and presidential candidates [207]. The principle of this method is to add random noise to the face photos at pixel levels, then to collect judgments on pairs of distorted photos (which picture looks more trustworthy, for instance) in terms of changes in specific impression traits, then to average the selected photos in the pair stimuli to visualize the holistic features underlying corresponding trait perception.

Similarly, the bubbles technique has been used to identify the diagnostic regions and spatial scales of trustworthiness and dominance. With the learned diagnostic regions and sensitivity to contrast, they were able to manipulate photos to increase or decrease perceived impressions. See Fig 2.4 for details.

2.3 Social psychological factors related to impression formation

Face perception is a dynamic and adaptive process. Impressions of a face are dependent on factors other than the face per se. Cognitive states and social context jointly affect evaluations of faces.

2.3.1 Cognitive states

The typicality of a face has been shown to correlate with the perceived trustworthiness [184] and attractiveness [144] positively.

Studies have shown that typicality can be shifted quickly by just a few minutes' adaptations to a set of weird or atypical faces[144, 198], and attractiveness ratings have been distorted toward the adjusted center of "typical" faces[144].

The mere increased exposure can increase liking of the face[214], which goes along with the theory that typicality and familiarity facilitate positive perception. Apart from prior exposure,

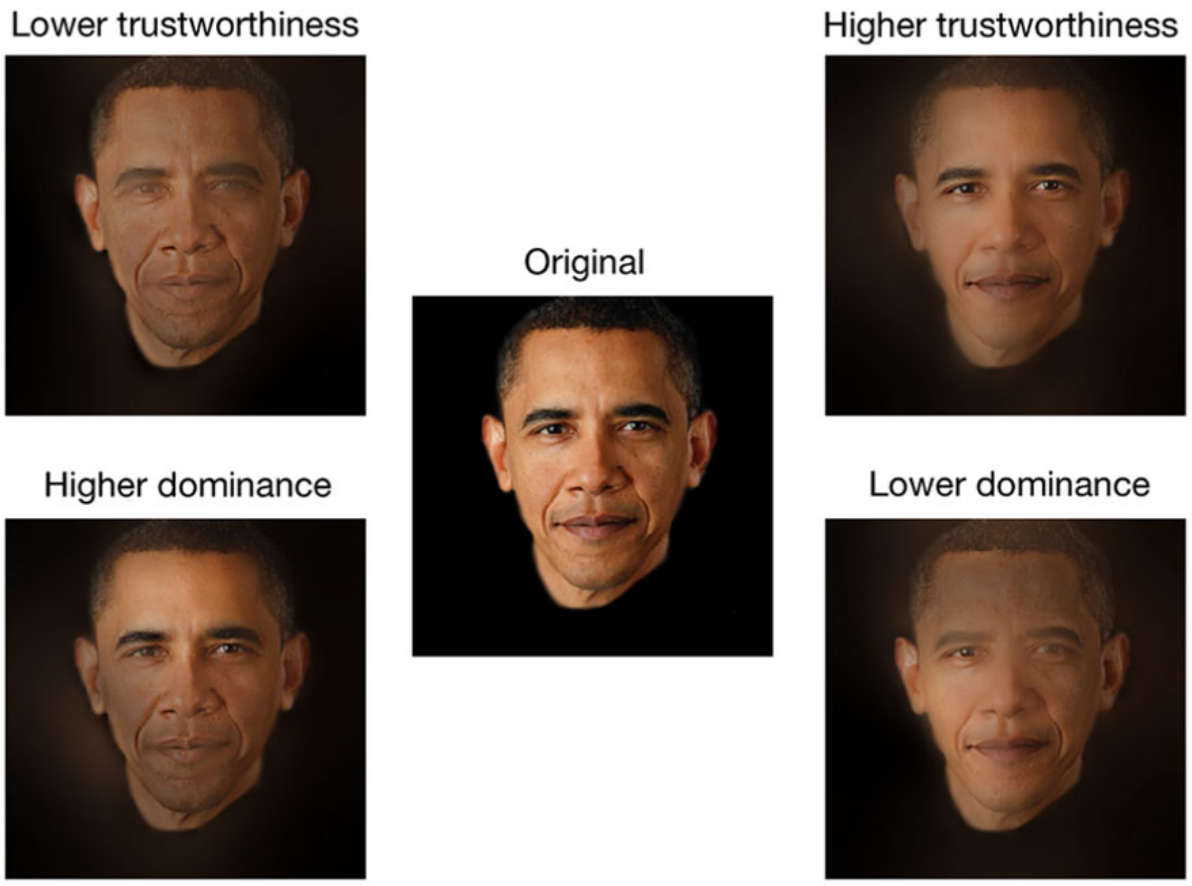


Figure 2.4. Manipulation of trustworthiness and dominance via contrast changes. Extracted from Robinson [147].

prior expectations about a person's personality can also intervene with the visual processing of the face.

An anecdotal example raised by Hassin and Trope in[67] is that knowing that Einstein is intelligent is likely to change people's visual perception of his forehead to be more significant. Hassin and Trope have found that two processes "reading from faces" (with visual cues), and "reading into faces" (prior expectation of a person) are two sides of the same coin, and both processes interact during impression formation. Transient, incidental association with faces also plays a role in changing impressions. When participants are asked to withhold a response to a face in a cognitively demanding task, the face is perceived as less trustworthy [41]. In a similar vein, when participants need to conduct a categorization task before rating a face, the more difficult the categorization task is, the less attractive the face that follows looks[140]. This phenomenon of "mind-at-ease-put-a-smile-on-face" leads to the processing fluency theory, which argues that beauty is grounded in the processing experiences of the perceiver[140]. This theory also explains why prototypical faces are perceived as attractive since they are easy to process[203].

Face-resemblance is another factor that has been noticed by psychologists to facilitate positive impressions. DeBruine and her colleagues found that faces with a resemblance to self increase the attractiveness and trustworthiness of a face[37, 38]. More interestingly, novel faces that resemble the faces of significant others are rated similarly to the significant others[65].

2.3.2 Social context

In social psychology, "stereotype" refers to an over-generalized belief about a particular category of people. To be more specific, the belief encompasses the group's typical roles, behavior, occupations, physical appearances, personality characteristics, attributes, and behaviors [73]. By definition, stereotypes are embedded in our first impressions of an individual based on our prior stored knowledge of the social category the person belongs to (gender group, race group, age group, and religious group, etc.) For instance, females are perceived as less aggressive,

warmer, and more caring compared with males [11].

In 1954, Secord examined how social category-based stereotypes affect personality perception[161]. Implicit bias test[62] has been widely used to capture an individual's prejudice and stereotype toward various social groups, and it is naturally linked to face perception. Research has shown that implicit racial bias predicts decisions based on faces in trust games [170]. When the context changes, the evaluation of racially ambiguous faces alter accordingly [80].

One influential theoretical work in social stereotype research is the "stereotype content model" (SCM), which was developed by Fiske and her colleagues[52, 53]. This model characterizes the universal impressions of social cognition for different social groups, see Fig2.5 (e.g., students, homeless, Asian, etc.) The primary dimensions proposed in the SCM are competence and warmth. For instance, "Asians" and "rich" people are perceived as high in competence and low in warmth; the homeless are perceived as low in both warmth and competence; the elderly are perceived as high in warmth but low in competence. Interestingly, the key dimensions in this model have a striking correspondence with the two dimensions discovered in facial first impression studies[181], dominance, and trustworthiness. "Warmth" and "trustworthiness" relate to the intentionality of the other person, whereas "competence" and "dominance" relate to the perceived capability of the person. The resemblance of the dimensions underlying facial impressions from visual cues and stereotypes of social groups (abstract knowledge) suggests that these dimensions are fundamental to the social perception of people.

Another influential work by Freeman and Ambady [54] proposed a dynamic interactive theory of person construal. The model incorporates the top-down influence of social categories, stereotypes, high-level cognitive states, and bottom-up processing of visual, vocal, and bodily cues (see Figure 2.6). Although in this review, we only focus on faces, this model offers a theoretical framework that stitches the top-down and bottom-up processes together. Visual and other sensory inputs feed into the category level and then activate the stereotypes associated with each category. Higher-order cognitive states such as priming, expectation, motivation, cognitive loads, or attention have a bi-directional connection with the stereotypes and nodes at

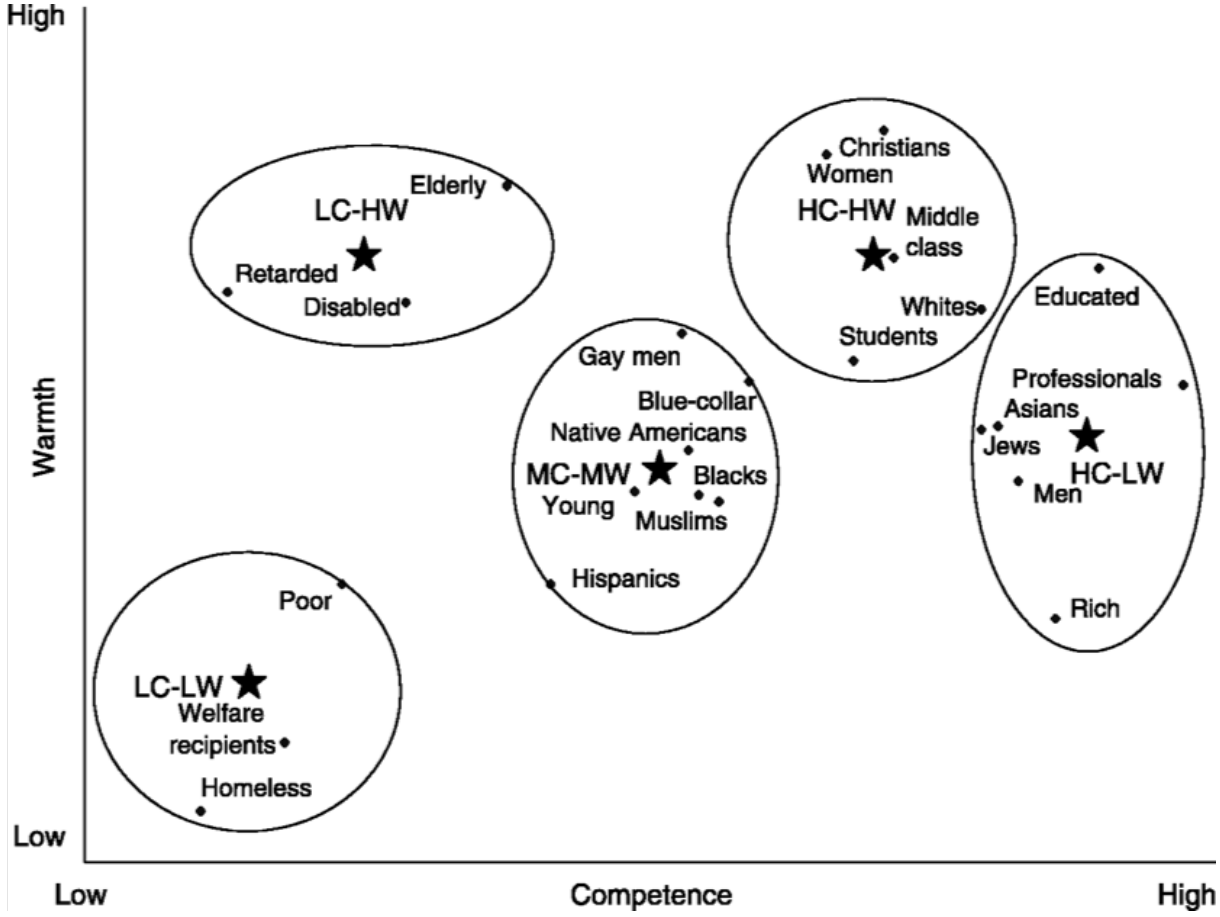


Figure 2.5. Stereotype content model. Extracted from [53]

the category level. This theoretical framework is supported by a number of empirical studies in social categorization [82, 56].

2.4 Computational modeling of facial impressions

2.4.1 Predictive models

Most face processing models in computer vision focus on the objective aspects of faces, such as face identity[190, 135], face alignment[204], and face attribute prediction (age, gender, pose, ethnicity, expression)[113]. There are relatively fewer works on predicting the subjective properties of faces, such as facial first impressions. Most work has been done on predicting facial attractiveness, both with traditional computer vision techniques and with the more advanced deep learning techniques. [48] used landmark-based geometric ratio-related facial features to build an attractiveness predictor (0.65 correlation with human ratings, face database size = 184). Amit Kagian and his colleagues have used a combination of landmark-derived features along with global features to obtain a high correlation with human group averages on facial attractiveness (0.82 Pearson correlation, face database size=91)[85]. Traditional computer vision features such as SIFT, HoG, and Gabor filters have been blended to predict the relative ranking of facial attractiveness in [4] (rank correlation 0.63, face database size=200). [148] incorporated collaborative filtering techniques with visual features extracted from pretrained VGG networks [162] to achieve individual-level prediction of facial attractiveness[148] (correlation 0.671 on female face queries, database size = 13,000). [122] build a model based on a pretrained VGG network to predict trustworthiness, dominance, and IQ in faces (R2 values on trustworthiness, dominance, and IQ are 0.5687, 0.4601, 0.3548 respectively, face database size=6000). End-to-end neural networks were applied to predict facial attractiveness in 2010[61] (correlation 0.458, face database size = 256, young female faces only). Another work collected a dataset consisting of 1000 Caucasian faces with ratings on 16 social traits, then built a neural network based on fiducial points and achieved rank correlation of 0.90 on approachability, 0.70 on

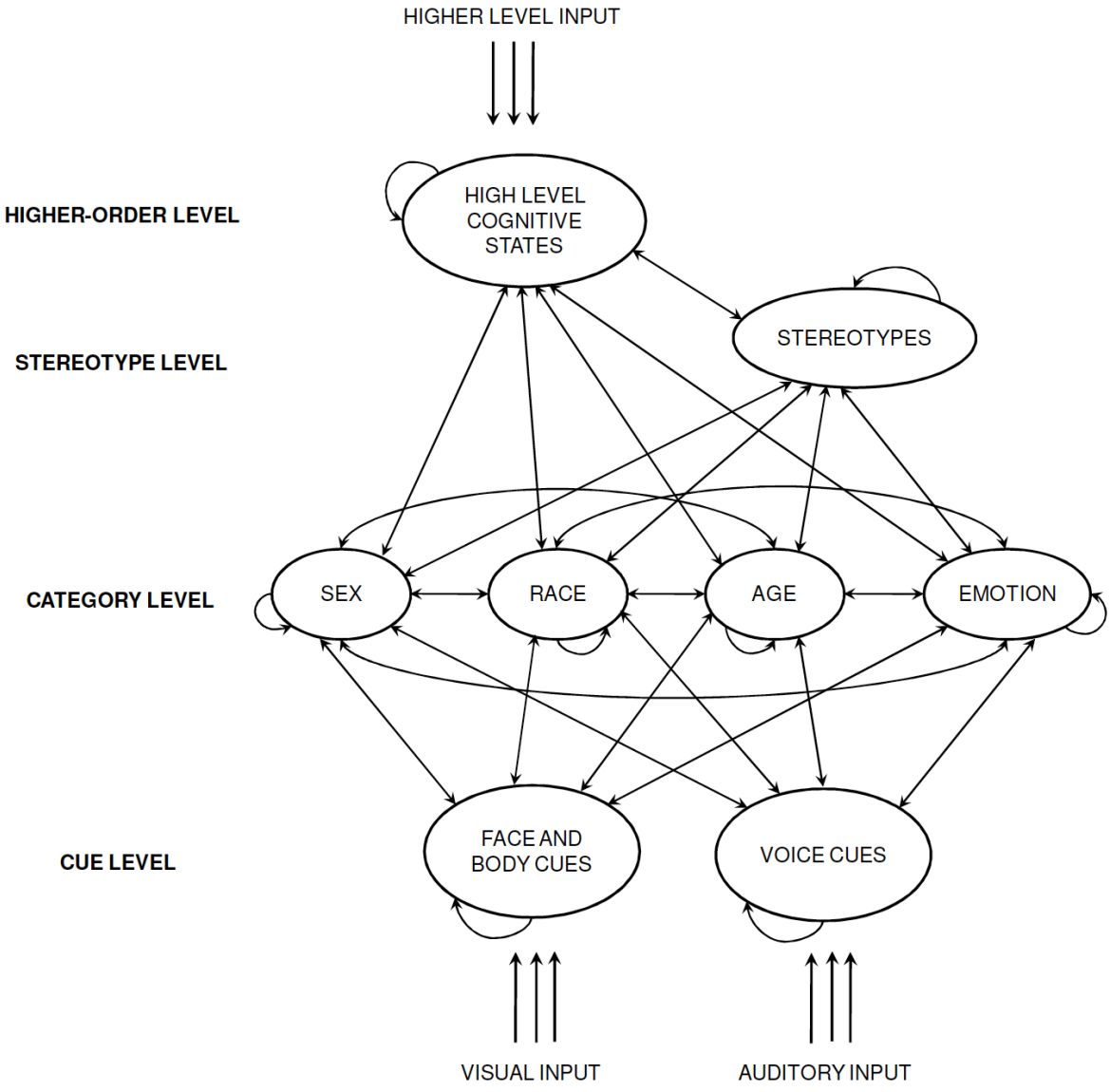


Figure 2.6. Dynamic interactive model of person construal. Extracted from Freeman and Ambady [54].

youthful-attractiveness, and 0.67 on dominance[194]. [63] has proposed a method based on Active Appearance Model (AAM)[30] to extract visual features from faces then built regressors to predict the social attributes collected in US 10 dataset[63]. Our own work on predicting 40 social attributes based on state-of-the-art pre-trained neural network models will be presented in Chapter [168].

Previous papers have achieved correlations with human performance between 0.458 to 0.82 in attractiveness predictions, depending on the dataset and method used. However, to date, there is no standard dataset that has been used to compare these approaches. Earlier studies employ datasets with relatively small numbers of faces (a few hundred), and most face datasets use young Caucasian faces only, as pointed out by [102]. In contrast, US 10K dataset[89, 8] consists of 2,222 high-quality color images that vary in ethnicity, gender, age, and expression, with ratings on 40 attributes. This dataset is smaller than two of the ones mentioned above. The first is collected from howhot.io, an online dating website[148] and contains 13,000 face images, but that work focused on personalized prediction of facial attractiveness, rather than average ratings. There are only binary choices (hot or not) indicating implicit preference of facial attractiveness. The second one is collected from testmybrain.com, containing 6,000 grey-scale face images[122], and includes just three social features: dominance, IQ, and trustworthiness.

2.4.2 Modification models

In the computer vision literature, there are two related tasks: image modification [79] and image generation [59, 189]. In image modification tasks, the aim is to edit certain aspects of an image while keeping other aspects unchanged, e.g., changing the skin tone of a face without changing any other property of it, which is often called image translation. Another type is to modify a continuous trait of an image, such as the age of a face, or the trustworthiness level of a face.

The task of image generation is to generate an image given certain conditions, e.g., given a category of an object. Both problems require capturing the underlying representation of the

target.

Various approaches have been developed to tackle these problems: template-based models[30, 182, 63], restricted Boltzmann machines, discriminative approaches[189] and autoencoders[74, 31]. Recently, generative methods such as GANs and VAEs have been proposed to achieve realistic performance.

Initially, much of this work required paired examples or was done with hand-tailored methods [88, 99, 71, 90, 43]. This method can be time-consuming and can struggle with generalization. More general approaches have been introduced, but only for the image to image translation. Some of these newer methods can modify continuous traits, such as the apparent age of a face, but they can only achieve this by breaking images into binary categories such as "young" and "old" or multiple age groups, such as 0-18, 19-29, etc. [191, 58, 111, 5].

Another level of complexity is introduced when dealing with unpaired data [218]. That is, given unrelated images from different domains, translate an image from one class to another. Without pairs of images in each domain, the problem cannot be solved as merely mapping an input image to its paired output image.

The unpaired image to image translation task has been tackled as a latent space traversal problem. For example, [191] uses deep features from a pretrained classifier to form a latent space in which different classes of images are loosely clustered. Then, by projecting an image into the latent space, interpolating the embedding towards the mean of a different class, and projecting back into the image space, class modifications can be achieved [191, 58]. However, these methods require multiple steps and use hand-tailored linear interpolations through a latent space. These methods take more time to design and do not generalize well to new domains. A more desirable method of modifying an image would do so based purely on learned weights, using a single forward pass of a neural network.

Over the past four years, generative adversarial networks [59] have become the most popular tool for the image to image translation tasks, allowing for the image modification to be fully controlled by the generator network in one forward pass. They use adversarial training to

ensure results are believable and often use cycle training to ensure that identifying traits from the original image are preserved [218, 78, 112, 26, 32, 111, 101, 138]. However, by definition, GANs require at least two networks to be trained. These networks also need to remain in equilibrium throughout the training process to prevent modal collapse [15]. A single-network image modification method would be better since it would have fewer parameters and train in a faster and more straightforward manner.

None of these methods are built to work on continuous trait modification. They currently only use binary classes (old vs. young, or beard vs. clean shave). While it is possible to interpolate between these binary classes linearly, traits may not scale linearly between these classifications, e.g., aging between 10 to 15 years old would require a different change than aging from 30 to 35 years old. As mentioned above, existing techniques either use hand-tailored methods, rely on multiple steps like projection and linear interpolation, or require two networks for training.

2.5 Cultural comparison of first impressions

Another dimension needed to look at social impression formation is culture.

As Hofstede put it [75], "culture consists of patterned ways of thinking, feeling and reacting, acquired and transmitted mainly by symbols. The essential core of culture consists of traditional ideas and their attached values." East Asian people have been characterized as being more collectivists, whereas Western people have been described as more individualistic[75, 133]. It is natural to ask whether cultural values will affect the way people form impressions of each other. There has been a large body of literature examining the similarities of face perception across cultures. Previous studies have shown that emotion recognition is in large the same across cultures[49, 156], although there are also some variations [35, 50]. In the domain of facial impressions, studies have found both similarities and specifications across cultures. In one study, researchers found US and Korean subjects have high agreement on babyfacedness[121]. Another

study found that black American, white American, and Korean participants agree highly on a range of impression traits (attractiveness, babyfacedness, honesty, naivete, submissiveness, and warmth) of same-race and other-race faces. A similar study examined the consistency in facial attractiveness and found high consensus in US perceivers and Taiwanese immigrants on facial attractiveness[34]. Albright and her colleagues found cross-cultural agreement in personality judgments[3] in China and the US. Empirical evidence demonstrated that Tsimane people who live in the remote Bolivian rain forest also share consensus in first impressions of faces (in terms of attractiveness, babyfacedness, health, intelligence, dominance, and sociability) with people in the US [213]. Rule and colleagues found the US and Japanese people show high agreement in their impressions of CEOs [153]. On the other hand, a comparison of perceivers in Japan and Israel found that culture-specific face typicality influences perceptions of trustworthiness[166].

Another line of studies compares the key dimensions of impression traits across cultures. Using principal component analysis, Todorov and his colleagues found two key dimensions underlying Caucasian faces[181] are trustworthiness and dominance. Research on social impressions of different social groups [33] found that warmth and competence are universal across all cultures. The similarity between key dimensions of facial impressions and social impressions of different social groups suggests that impression formation is a general social-psychological process and may have an evolutionary basis. Following this line, Sutherland and her colleagues have extended the model developed by Todorov using highly variable face photos and identified a third dimension: "youthful-attractiveness" in Caucasian faces [175]. They further broaden the same method to identify the cultural similarities and differences across perceivers and faces in China and the UK[173]. They found a common approachability and youthful-attractiveness dimensions across perceiver and face race, with some evidence of a third dimension akin to capability[174]. In a pre-registered study[83], a large group of researchers plans to evaluate Todorov's valence-dominance[181] model around the world (Africa, Asia, Australia, and New Zealand, Central America and Mexico, Eastern Europe, and the Middle East, the USA and Canada, Scandinavia, South America, the UK, and Western Europe) using 120 face images

consisting of Black, White, Asian and Latin faces.

2.6 Summary

The social perception of faces is a highly interdisciplinary subject, and it is an exciting and vital time to study the topic. A growing number of people are posting photos on-line and have on-line social interactions. Research results from psychology and social science have shown insights into the universal and unique aspects of impression formation. Advances in computer vision and machine learning techniques have given the power to researchers to generate and manipulate faces with unprecedented control and realism. Equipped with these theoretical frameworks, behavioral methodologies, and computational models, we are aiming at achieving a deeper understanding of how people process the social signals in faces and how cognitive functions interact with culture and decision making.

Chapter 3

Predicting human impressions of faces

3.1 Introduction

Humans are skilled in extracting a variety of information from faces: identity, age, gender, ethnicity, etc. Apart from these objective properties, humans also quickly form subjective impressions of faces at first glance[184], such as facial attractiveness[179], friendliness, trustworthiness[181], sociability[23] and dominance[123]. In spite of the subjective nature of these social judgments, there is often a consensus among humans in how they perceive attractiveness, trustworthiness, and dominance in faces [51, 48]. This consensus indicates that faces contain high-level visual cues for social inferences, therefore making it possible to model the inference process computationally. Social judgments of faces have a significant impact on social outcomes, ranging from electoral success to sentencing decisions[200, 132]. Therefore it is beneficial to understand the nature of social judgments of faces and be able to predict and modify these impressions.

3.2 Related Work

With a hypothesis-driven approach, psychologists have identified that high level visual features, such as the averageness of a face[100, 146, 144] and the symmetry of the face [159], can explain why certain faces look more attractive. Other global face features such as femininity, babyfaceness[14], typicality[167] drive different aspects of social impression perception (warmth,

honesty, submissiveness, dominance, etc.). Emotions also drive social impressions of faces. For instance, the amount of perceived anger or happiness will drive aggressiveness/dominance and trustworthiness perceptions, respectively. Using morphing and averaging methods, studies[175] have established that age serves an important role in social perception of attractiveness, trustworthiness and dominance.

Another body of studies focus on predicting social impressions from face photos in a data-driven fashion. Traditional computer vision algorithms have been used to predict facial attractiveness [61], trustworthiness [51], sociability, aggressiveness [123] and memorability [8, 89]. Simulated face models are used to predict and generate faces with specified impression scores[181, 182]. However, since the simulation model is trained on simulated faces, the computational model lacks the power to generalize to realistic face photos. One study used realistic faces, but its database is small and lacks diversity in expressions and ethnicity[197]. Active Appearance Model has been deployed to model and predict human impressions on multiple traits[63].

Apart from perceptual cues, properties of the observer play a role in the social perception of faces. Social and cultural[55] attitudes, such as implicit race bias[170] , can also influence how faces are perceived. Perceptual fluency can affect how people perceive the face[203, 140]. This has been described as the "a mind at ease puts a smile on the face" theory[202]. Faces that resemble the observer's own face are perceived as more positive and trustworthy[36].

Despite the complexity of first impression formation, given that to a large extent, humans share consensus on some of these impression judgments, this research focuses on predicting humans' shared consensus of impression attributes of faces. Motivated by the success of deep learning in modeling the objective properties of faces[178, 171, 66], we use deep learning models to be the basis for learning to predict first impressions of realistic faces.

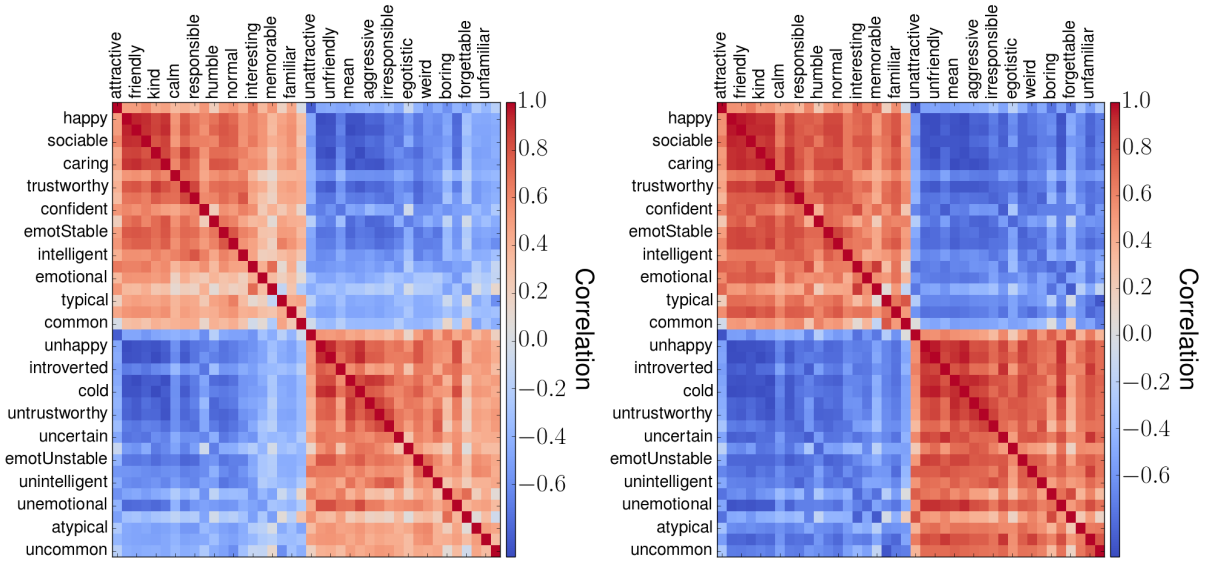
3.3 Dataset

Our model focus on a set of impressions traits identified by psychologists as being fundamental features of first impressions in many contexts.

To capture a broad range of human social impressions of faces, we use US Adult 10K database [8] consisting of 2,222 face images and annotations for 40 social attributes. Each attribute is rated on a scale of 1-9 by 15 subjects. We take the average rating from all raters as a collective estimation of human judgment for the social features of each face.

The 40 social attributes consist of 20 pairs of related traits: (attractive, unattractive), (happy, unhappy), (friendly, unfriendly), etc. Some of these traits are highly correlated and predictable from others, especially within the trait pairs. To understand the human-perceived correlations between these traits, we compute the Spearman's rank correlation between the average human ratings of every pair of social features and show their correlations in a heatmap (Figure 3.1(a)). We order traits in the map based on similarity and positive or negative connotation. From the figure, we see that negative social features such as untrustworthy, aggressive, cold, introverted, and irresponsible form a correlated block. Likewise, the most positive features such as attractive, sociable, caring, friendly, happy, intelligent, interesting, and confident are highly correlated with each other. Although we choose 20 pairs of opposite features, they are not completely complementary and redundant. Principal Component Analysis of the covariance matrix shows that it takes 24 principal components to cover 95% of the variance.

After averaging human ratings, each face receives a continuous score from 1 to 9 in all social dimensions. We model these social scores with a regression model. We propose a ridge regression model on either features from deep convolutional neural networks (CNN) or traditional face geometry based features, and present results from both feature sets. Such visual features are usually high-dimensional, so we first perform Principal Component Analysis (PCA) on the extracted features of the training set to reduce dimensionality. The PCA dimensionality is chosen by cross-validation on a validation set, separately for each trait. The PCA weights are



(a) (b)

Figure 3.1. Correlation heatmaps among social features. (a): human; (b): CNN-based model.

saved and further used in fine-tuning our CNN-regression model.

3.3.1 Regression on Geometric Features

Past studies have found that facial attractiveness can be inferred from the geometric ratios and configurations of a face [48, 85]. We suggest that other social attributes can also be inferred from geometric features. We compute 29 geometric features based on definitions described in [117], and further extract a 'smoothness' feature and 'skin color' feature according to the procedure in [48, 85]. The smoothness of a face was evaluated by applying a Canny edge detector to regions from the cheek and forehead areas [48]. The more edges detected, the less smooth the skin is. The regions we chose to compute smoothness and skin color are highlighted in the right subplot of Figure 3.2. The skin color feature is extracted from the same region as smoothness, converted from RGB to HSV. However, regressing on these handcrafted features alone is not enough to capture the richness of geometric details in a face. We therefore use a computer vision library (dlib, C++) to automatically label 68 face landmarks (see Figure 3.2) for each face, and then compute distances and slopes between any two landmarks. Combining 29 handcrafted

geometric features, smoothness, color and the distance-slope features, we obtain 4592 features in total. Since the features are highly correlated, we apply PCA to reduce dimensionality. Again, the PCA dimensionality is chosen by cross-validating on the hold out set separately for each facial attribute. Then a ridge regression model is applied to predict social attribute ratings of a face. The hyper-parameter of ridge regression is selected by leave-one-out validation within the training set.

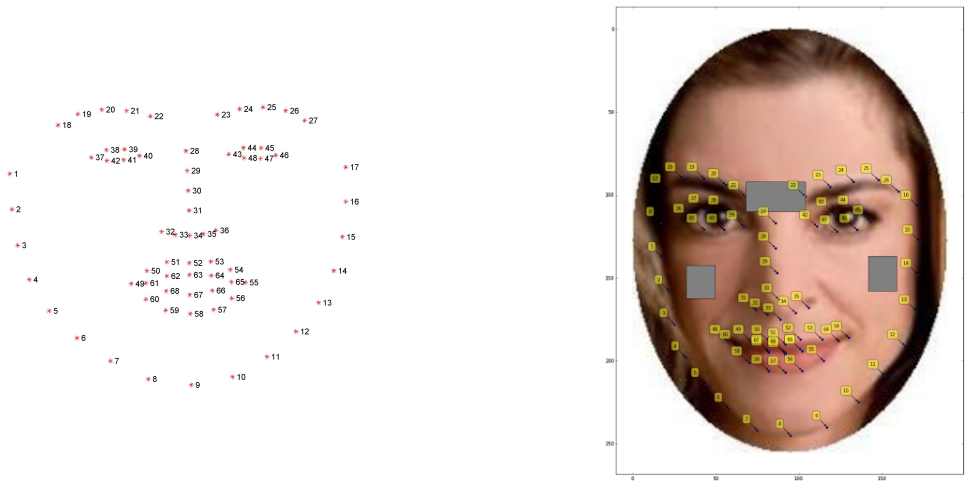


Figure 3.2. 68 face landmarks labeled by dlib software automatically. The gray regions are used for computing smoothness and skin color.

3.3.2 Regression on CNN Features

Previous studies have shown that pretrained deep learning models can provide feature representations versatile for related tasks. We therefore extract image features from pretrained neural networks, choosing from six architectures with different original training goals: (1) VGG16, trained for object recognition [162], (2) VGG-Face, trained for face identification [162], (3) AlexNet, trained for object classification [95], (4) Inception from Google, trained for object recognition [177], (5) a shallow Siamese neural network that we train from scratch to cluster faces by identity, (6) a state of the art VGG-derived network (Face-LandmarkNN) trained for the face landmark localization task.

To find the best CNN features among the six networks, we first find the best-performing

feature layers of each network in the ridge regression prediction task. Before the ridge regression, we perform PCA and pick the PCA dimensionality that gives best results on the validation set. Then, we compare the results among networks to select the best features overall.

3.3.3 Results

After comparing all 6 networks, we find that the conv5_2 layer of VGG16 (trained for object classification) lead to the best results. This set of features significantly outperforms the three networks trained solely on faces, while also slightly outperforming AlexNet and Inception networks. These best-performing CNN features also exceed the prediction correlation of the geometric features in most attributes. Figure 3.3 compares prediction performance of the CNN model and the geometric feature model.

We speculate that the poor performance from the face recognition networks can be attributed to their optimization for specific facial tasks. Learning face landmark configurations and differences between faces that define identity may not correlate well with the task at hand, which looks for commonalities behind certain social features beyond identity. The landmark networks should presumably give results similar to the geometric features, but did not learn features corresponding to all of the features we manually extracted.

We also try fine-tuning the best performing CNN model with back propagation but do not observe further improvement in performance. Hence our reported results are without fine-tuning.

To evaluate model performance, we did a random train/validation/test split 50 times, with a ratio of 64/16/20 respectively. The prediction performance of our model is evaluated using Pearson's correlation with the average human ratings on the test set. For each social attribute, we also compute human group consistency as an index of the strength of learning signal.

Among the social attributes, human subjects agree most about 'happy' and disagree most about 'unfamiliar'. For both regression models (CNN based regression and geometric feature based regression), model performance grows as the consensus on a social trait increases.

Since a change in expression would produce a change in landmark locations, it is

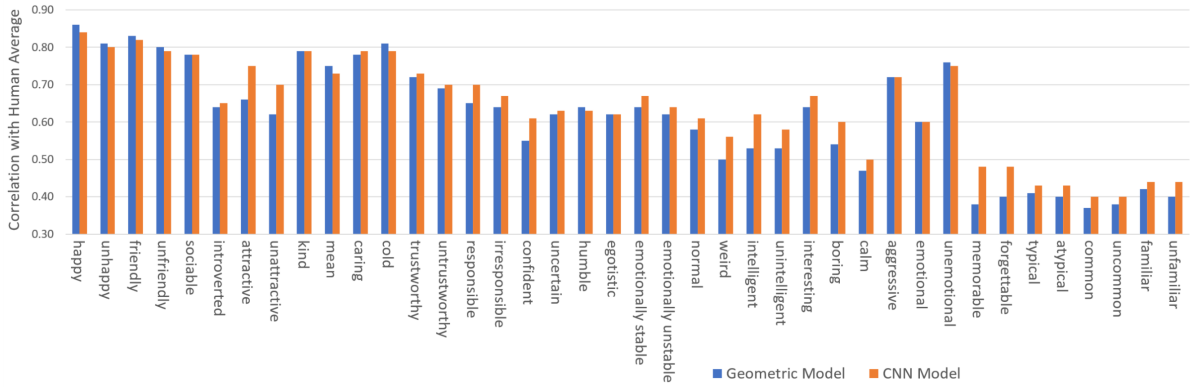


Figure 3.3. Model comparison on 40 social features.

not surprising that landmark-based geometric features achieve comparable or slightly higher correlation with the CNN model when predicting social attributes which are highly related to expressions (such as 'happy', 'unhappy', 'cold' and 'friendly' etc). For other social attributes, the CNN model performs better, by about 0.04 higher in correlation on average. This implies that CNN features encode much more information than landmark-based features. It is useful to visualize such features to understand what aspects make them powerful enough to predict social attributes.

3.3.4 Evaluating Against Human Consensus

An important gauge of model success is quantitative comparison between the subjective social features predicted by our best performing model and those perceived by humans. We take our model predictions, compute the Spearman correlation between every pair of traits, and display them in a heatmap (see Figure 3.1 (b)). The resulting heatmap shares similar patterns with the figure generated from average human ratings (see the left panel in Figure 3.1). Pearson Correlation between the upper triangle of the two similarity matrices (human and model prediction) is 0.9836. This suggests that our model successfully preserves human-perceived relationships between traits.

Since these social impressions are subjective ratings, it is informative to examine the

extent with which people agree with each other on these judgments. To calculate human group consistency, we perform the following procedure 50 times for each attribute and then average the results: (1) For each face, we randomly split the 15 raters into two groups of 7 and 8. (Note: The raters assigned to each face are generally different sets). (2) We calculate the two groups' average ratings for each face, obtaining two vectors of length 2,222 (the number of faces in the dataset). (3) Finally, we calculate the Pearson correlation between the two vectors. We find that human agreements covary with model performance and observe an extremely high correlation, as illustrated in Figure 3.4.

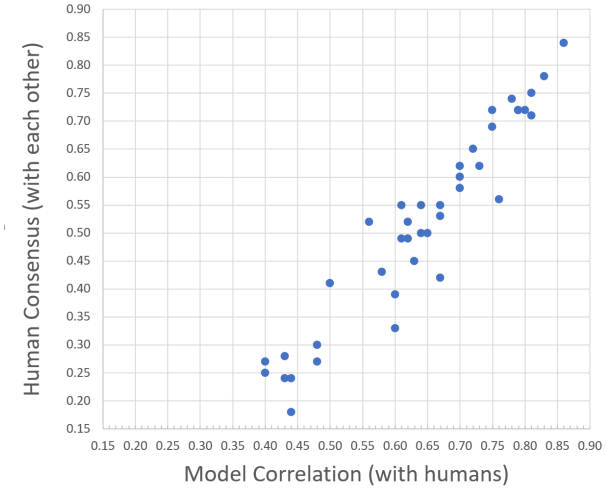


Figure 3.4. Human within group consistency vs. model's correlation with human average. Pearson correlation $\rho = 0.98$, $p < 10^{-5}$

3.4 Feature Visualization

Here, we visualize features from our model which are important for social perceptions. We tried on two methods.

3.4.1 Deconvolution

We choose facial attractiveness as an example, but the same method can be applied to the other social features. To identify visual features that ignite attractiveness perception, we find

the top 9 units of highest influence on attractiveness at conv5_2 as follows. First, we compute a product of three terms: (1) A unit's activation from conv5_2, (2) that unit's weight to the following fc_PCA layer, (3) the fc_PCA unit's weight to the output unit. We then sort all conv5_2 units' average products of these three terms and identify the top 9 neurons that contribute to the output neuron for the corresponding social feature. Then we employ the method described in [206, 215] to find top-9 input images that cause high activations in each of the top-9 conv5_2 neurons. Also we use deconvolution to create an image of the features activating that unit for each face, with varying levels of success.

Figure 3.5 captures the features that are important for predicting the attractiveness of a face. The feature importance descends from left to right and top to bottom. The important features identified by our model are related to eyes, hair with bangs, high nose-bridges, high cheeks, dark eyebrows, strong commanding jawlines, chins, and red lips. Note that among the 9 cropped input image patches, not all the faces are perceived as attractive overall; despite having a feature that contributes to attractiveness. An attractive face needs to activate more than one of these features in order to be considered attractive. This observation agrees with our intuition that attractiveness is a holistic judgment, requiring a combination of multiple features.

It also seems that several attractiveness features include relationships between different facial features. For example, while the first feature in the upper left of the figure emphasizes the eye, it also includes the nose. This is also true of the upper right feature. Additionally, smiling is important in perceived attractiveness, as emphasized by the feature in the lower left of the figure.

3.4.2 Global Average Pooling

Another method to visualize the convolutional neural network is via global average pooling on the convolutional layers[217]. This method allows us to identify the exact image regions diagnostic of each individual facial impression trait, exposing the implicit attention of the neural network on an image. To be specific, we use VGG16's network structure[162], but throw

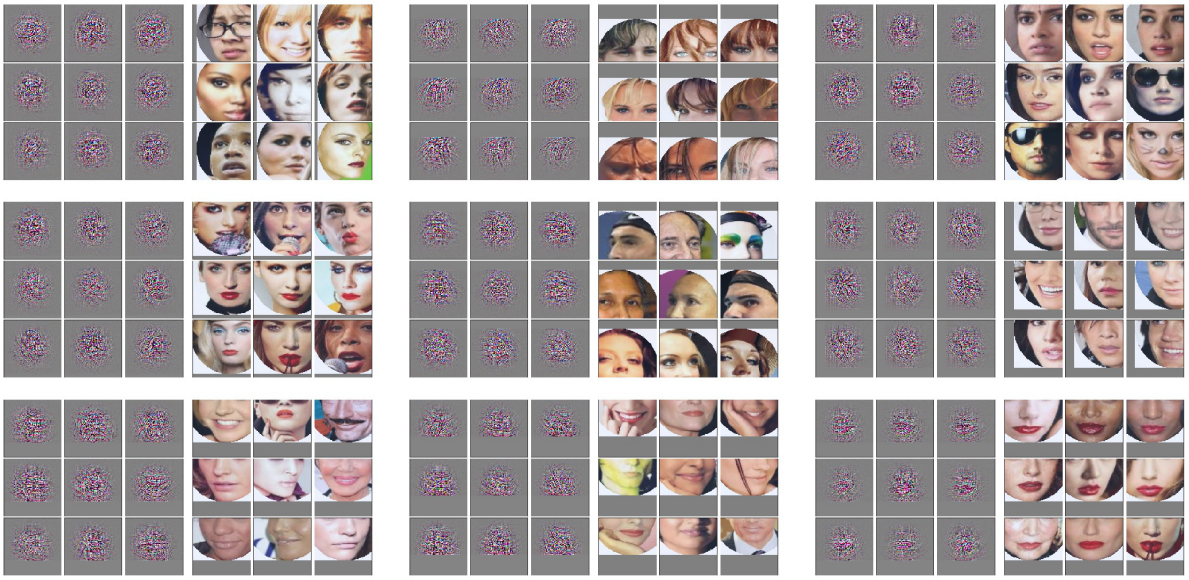


Figure 3.5. Visualization of features in the pretrained-VGG16 regression network. For conv5_2 layer, we show the top 9 activation of the top 9 neurons that maximally activate the attractiveness neuron across the training data, projected down to pixel space.

away the fully connected layers. Just after the last convolutional layer (conv5_3), we perform global average pooling on the convolutional feature maps and these features for a fully-connected layer to produce regression results to the 40 social attributes. For a given image, let $f_k(x,y)$ represents the activation of unit k in the last convolutional layer at spatial location (x,y). Then, for unit k, the result of performing global average pooling is F_k,

$$F_k = \Sigma_{x,y} f_k^t \tag{3.1}$$

Thus, for a given trait t, the predicted score S_t is $\Sigma_k w_k^t F_k + b^t$, where b^t is the bias term for trait t, whereas w_k^t indicates the importance of F_k for trait t. We define M_t s the impression activation map for trait t, where each spatial element is given by:

$$M_t(x,y) = \Sigma w_k^t f_k(x,y) \tag{3.2}$$

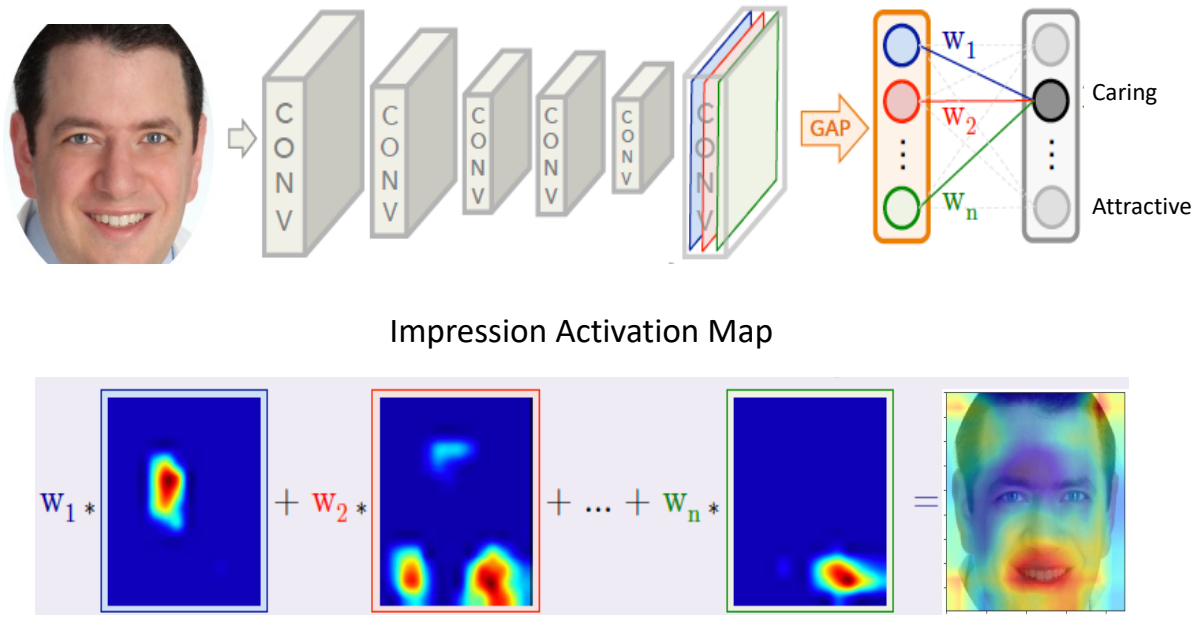


Figure 3.6. Impression activation map: the predicted score is mapped back to the previous convolutional layer to generate the impression activation maps (IAM). The IAM highlights the trait-specific diagnostic regions.

So $S_t = \Sigma_{x,y} M_t(x,y)$, and $M_t(x,y)$ indicates the importance of the activation at spatial grid (x,y) leading to the regression score of an image to trait t.

Compared with our previous CNN architecture, there is a drop in prediction performance, as can be seen in table 3.1 , but the similar method can be applied to the original model to generate similar heatmaps. In Fig 3.7, we show two examples of the IAMs output using the methods mentioned above. We can see that the mouth areas and the smiles are captured in the happiness detection maps. The attractiveness activation maps are more holistic and highlight both the mouth region and the forehead regions.

3.5 Conclusion

In this work, we address the problem of predicting first impressions on face, and have shown that a deep network can be used to predict human social perception of faces. We show that empirically, our method is highly effective, achieving high correlation with the average

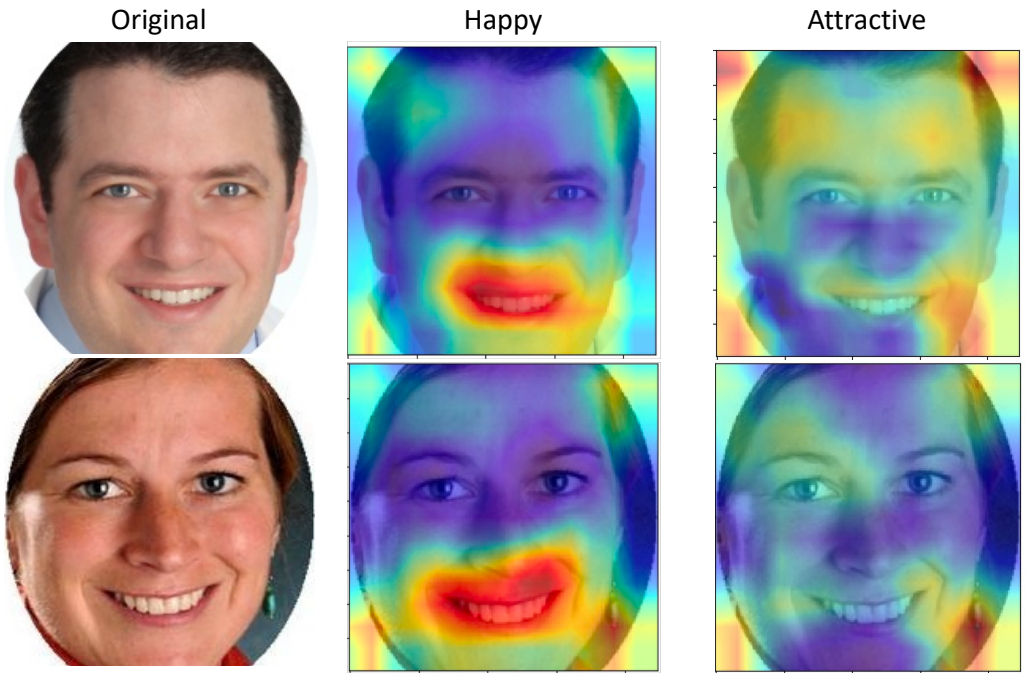


Figure 3.7. Examples of impression activation maps in two photos, the first column is the original face photo, the second and third column are the corresponding happy and attractive activation maps. The maps highlight the regions used for trait prediction.

human ratings. As far as we know, this is the widest exploration of social judgment predictions, showing human-like perceptions on 40 social dimensions. The simplicity of our method makes it portable and easy to apply to any new type of social attribute, such as the "corporativeability" of human faces.

Reflecting previous work in recognizing facial expressions, where happiness is the easiest to recognize, our highest correlation is on the happy feature. However, previous work in this area tends to classify a face as happy or not, rather than the degree of rated happiness. By predicting this as a continuous value, rather than categorical data, the subjective nature of human judgment is modeled smoothly, along with the subjective face trait landscape.

We find that, for attributes which are recognized via facial actions, such as happy, unhappy, or aggressive (probably associated with anger) or lack of facial action, such as cold or unemotional, a simple regression model based on the placement of facial landmarks works well, although the deep network performs nearly as well.

Of greater significance is our model's correlations with human judgments for traits such as trustworthiness, responsibleness, confidence, and intelligence, which correspond to more static features of the face. In this area, the deep network, which responds to facial textures and shape, has superior performance. While these judgments do not correspond to the traditional notion of "ground truth", they are descriptions for which humans have a fair amount of agreement, suggesting the presence of a signal to be recognized.

Furthermore, we have shown, yet again, that a machine can recognize attractiveness. For this dataset, our deep network correlates with average human ratings at 0.75. This provides a new benchmark for this dataset. This is one of a few areas where the deep network significantly outperforms the geometric features, as skin texture is likely to matter.

Many of these features are redundant. For example, friendly and happy are highly correlated (see Figure 3.1, and the red block indexed by happy and friendly). Similarly, aggressive and mean are highly correlated, which presumably requires *not* smiling. Meanwhile, it is also noteworthy that some traits considered to be "opposite" in this list are not simply the inverse of one another. For example, there is a large difference in human agreements on "sociable" (0.74) versus "introverted" (0.50), suggesting they are not opposites.

We also examined some of the features from the deep network. It is notable that these are difficult to verbalize, which is quite different from geometric features.

These results are significant for the field of social robotics. While a robot should not purely judge a human on appearance, much of human interaction is dictated by the underlying fabric of social impressions. Thus, it is important for a robot to be aware of this subjective social fabric, opening the door to useful knowledge such as whether humans might judge a person to be trustworthy. These judgments may happen subconsciously for humans, while a robot can be more objective, predicting these judgments and objectively choosing when to consider them in a decision. A robot need not treat an attractive or unattractive person differently for its own purposes, but this knowledge could affect how interactions are made for the sake of the human, knowing in advance how that person may feel that they fit into the social landscape.

Expansions on this work may include investigating the image properties that determine high level social features, beyond the attractiveness features. Additionally, social trait prediction may benefit from a single model with a shared representation, while this paper approaches each attribute as a separate regression task.

For future work, we aim to develop a generative model which can automatically modify a face's attributes (either objective or subjective) while preserving its realism and identity. Practically speaking, such a model could improve a face's perceived social features in positive ways (e.g. make a face look more sociable, trustworthy). More importantly, it would enable psychologists to quantify human biases during the formation of social impression in a precise and systematic manner. Psychologists could generate variants of a real face differing in age, gender, race, and explore how various factors separately and jointly affect the social impressions of faces.

3.6 Acknowledgement

Chapter 3 (predictive model), in part, is a reprint of the material as it appears in Proceedings of the 39th Annual Conference of the Cognitive Science Society (COGSCI'17). (Amanda Song, Linjie Li, Chad Atalla, Garrison Cottrell. "Learning to see people like people: Predicting social impressions of faces"). The dissertation author was the primary investigator and author of this paper.

Table 3.1. Human agreement on 40 social traits (measured by split-half rank correlation)/ Global Average Pooling-regression method's performance (measured by rank correlation with human averages on the test set) / CNN-PCA based regression method's performance.

Social Attribute	Human Agreement	GAP Model	CNN-PCA Model
happy	0.84	0.77	0.84
unhappy	0.75	0.72	0.80
friendly	0.78	0.74	0.82
unfriendly	0.72	0.71	0.79
sociable	0.74	0.65	0.78
introverted	0.50	0.64	0.65
attractive	0.72	0.71	0.75
unattractive	0.62	0.65	0.70
kind	0.72	0.72	0.79
mean	0.69	0.62	0.73
caring	0.72	0.66	0.79
cold	0.71	0.77	0.79
trustworthy	0.62	0.63	0.73
untrustworthy	0.60	0.57	0.70
responsible	0.58	0.55	0.70
irresponsible	0.55	0.50	0.67
confident	0.55	0.58	0.61
uncertain	0.45	0.51	0.63
humble	0.55	0.46	0.63
egotistic	0.52	0.57	0.62
emotionally stable	0.53	0.50	0.67
emotionally unstable	0.50	0.55	0.64
normal	0.49	0.52	0.61
weird	0.52	0.58	0.56
intelligent	0.49	0.40	0.62
unintelligent	0.43	0.39	0.58
interesting	0.42	0.47	0.67
boring	0.39	0.48	0.60
calm	0.41	0.35	0.50
aggressive	0.65	0.64	0.72
emotional	0.33	0.55	0.60
unemotional	0.56	0.71	0.75
memorable	0.30	0.37	0.48
forgettable	0.27	0.33	0.48
typical	0.28	0.33	0.43
atypical	0.24	0.32	0.43
common	0.25	0.30	0.40
uncommon	0.27	0.32	0.40
familiar	0.24	0.37	0.44
unfamiliar	0.18	0.31	0.44

Chapter 4

Modify faces with ModifAE

4.1 Introduction

Image modification is a challenging task that requires changing some aspects of an image while keeping others static. These changes can include modifying the class of an image, often called image to image translation [78], or they can refer to modifying a continuous trait in an image, which we will refer to as continuous image modification. These image modification tasks are often performed on datasets like CelebA because human faces have many easily recognizable traits [114].

Initially, much of this work required paired examples or was done with hand-tailored methods [88, 99, 71, 90, 43]. This can be time consuming and can struggle with generalization. More general approaches have been introduced, but only for image to image translation. Some of these newer methods can modify continuous traits, such as the apparent age of a face, but they can only achieve this by breaking images into binary categories such as "young" and "old" or multiple age groups, such as 0-18 years-old group, 19-29-years-old group, etc. [191, 58, 111, 5].

Another level of complexity is introduced when dealing with unpaired data [218]. That is, given unrelated images from different domains, translate an image from one class to another. Without pairs of images in each domain, the problem cannot be solved as a simply mapping an input image to its paired output image.

The unpaired image to image translation task has been been tackled as a latent space

traversal problem. For example, [191] use deep features from a pretrained classifier to form a latent space in which different classes of images are loosely clustered. Then, by projecting an image into the latent space, interpolating the embedding towards the mean of a different class, and projecting back into the image space, class modifications can be achieved [191, 58]. However, these methods require multiple steps and use hand-tailored linear interpolations through a latent space. These methods take more time to design and do not generalize well to new domains. A more desirable method of modifying an image would do so based purely on learned weights, using a single forward pass of a neural network.

Over the past four years, generative adversarial networks [59] have become the most popular tool for the image to image translation task, allowing for the image modification to be fully controlled by the generator network in one forward pass. They use adversarial training to ensure results are believable and often use cycle training to ensure that identifying traits from the original image are preserved [218, 78, 112, 26, 32, 111, 101, 138]. However, by definition, GANs require at least two networks to be trained. These networks also need to remain in equilibrium throughout the training process to prevent modal collapse [15]. A single-network image modification method would be better since it would have fewer parameters and train in a faster and simpler manner.

None of these methods are built to work on continuous trait modification. They currently only use binary classes (old vs. young, or beard vs. clean shave). While it is possible to linearly interpolate between these binary classes, traits may not scale linearly between these classifications, e.g. aging between 10 to 15 years old would require a different change than aging from 30 to 35 years old. As mentioned above, existing techniques either use hand-tailored methods, rely on multiple steps like projection and linear interpolation, or require two networks for training.

To fill this gap, we introduce ModifAE: a general image modification method which works for continuous traits, using only learned weights. To our knowledge, ModifAE is the first standalone neural network which can perform general image modification in a single forward

pass.

4.2 Related Work

4.2.1 Autoencoders

One method of generating efficient encodings of images is the autoencoder[74, 31]. These networks take in a high dimensional input, reduce to a lower dimensional representation, and then decode back to the original input. A bottleneck occurs in the middle of the autoencoder network, creating a latent space with high level features about the input. When the network is linear and has one hidden layer, they implement a version of PCA [134]. Additional hidden layers result in nonlinear encodings [74, 39]. More recently, variational autoencoders [93] have been introduced. These networks create more continuous latent spaces where linear interpolations can be performed smoothly.

Variational autoencoder

Variational autoencoder[94, 142] allows researchers to design complex generative models of data, yielding state-of-the-art performance in image generation and reinforcement learning.

A variational autoencoder consists of an encoder, a decoder, and a loss function.

An encoder is usually a neural network, and it takes a data point x as its input, and transform it into a lower dimensional latent representation z, which is referred to as a bottleneck of the representation.

The encoder is usually denoted as $q_\theta(z|x)$. The decoder is another neural network, which is denoted as $p_\phi(x|a)$. It takes the latent representation z as the input and reconstruct a data x from the network.

The loss function of the variational autoencoder is the negative log-likelihood with a regularizer.

$$l_i(\theta,\phi) = -E_{z\sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + KL(q_\theta(z|x_i)||p(z)$$

The first term is the reconstruction loss, and also the expected negative log-likelihood of the i-th data point. The expectation is take with respect to the the encoder's distribution $q_\theta(z|x)$over the representations, pushing the decoder to learn to reconstruct the data faithfully.

The second term is a K-L divergence regularizer between the encoder's distribution $q_\theta(z|x)$, the variational posterior, and $p(z)$. This divergence measures how much information is lost when using q to represent p. From a probabilistic point of view, variational inference approximates the posterior $p(z|x)$ with a family of distribution $q_\lambda(z|x)$. The variational parameter λ indexes the family of distributions. For instance, if q were Gaussian, then $\lambda_{x_i} = (\mu_{x_i}, \sigma^2_{x_i})$.

To track the variational inference, consider the following formula:

$ELBO(\lambda) = E_q[\log p(x,z)] - E_q[\log q_\lambda(z|x)]$.

We combine this term with the KL divergence and rewrite the evidence as: $\log p(x) = ELBO(\lambda) + KL(q_\lambda(z|x)||(p(z|x))$.

The Evidence Lower Bound allows us to make approximate posterior inference.

4.2.2 Deep Feature Interpolation

Some general methods of image to image translation are based on stepped linear traversals in a learned latent space. For example, Deep Feature Interpolation (DFI) [191] relies on linear interpolation of the latent representation of an input image. DFI involves training a deep convolutional neural network [163, 96] to classify images, then the network performs an image traversal based on a statically defined procedure, which is not learned by the network. Since DFI is designed to interpolate between means of class clusters, it cannot be used for continuous trait modification. Also, all modifications require a forward pass of the network, an interpolation in the latent space, then deconvolutions to arrive back at an image.

4.2.3 Conditional Generative Adversarial Networks

Conditional and Controllable GANs [124, 104, 205] made it possible to generate images conditioned on certain traits. The generator is trained with an image and trait input, then the

discriminator gives feedback based on whether the image appears real or fake and whether it appears to belong to the target class [104]. However, such networks can still struggle with preserving identifying traits of the original image.

[218] proposed **CycleGAN**, a model for image to image translation which addresses the challenge of maintaining identifying traits in the modified image. It enforces cycle consistency by modifying an image with a different target class, then modifying it back to its original classification. After the modification cycle, pixel-wise autoencoding loss is calculated for the original and reconstructed images. However, CycleGAN is incapable of modifying multiple traits with a single model.

Recently, **StarGAN** [26] was introduced and beat other methods on image to image translation. It uses cycle training to preserve identity traits and can also modify multiple classes in the same forward pass. [26] found that their model produced superior quality modifications when simultaneously supervising multiple traits, achieving state-of-the-art performance.

These networks can perform image to image translations on multiple traits in a single forward pass of the generator, while maintaining identifying traits of the original image. However, as mentioned in the introduction, they are currently all built to modify binary class traits. They also all require training multiple networks in equilibrium, which takes more parameters and more time. ModifAE offers a solution to both of these issues.

4.3 Methods

ModifAE is a single network, trained exclusively on an autoencoding task, that implicitly learns to modify perceived traits in images (illustrated in Figure 4.2). In this section, we discuss the collected dataset, ModifAE architecture, training procedure, and why ModifAE works.

4.3.1 Constructing a Continuous Trait Dataset

To train a generative model on continuous face traits, we need a large dataset. We use images from the CelebA dataset[114], which consists of over 200,000 images of celebrities

(including, ironically, B-list celebrities!). The images in CelebA are annotated with 40 facial traits, such as “mustache”, “eyeglasses”, “hat”, etc. Given that these are binary traits, rather than continuous ones, these labels are not appropriate for our purposes.

To generate continuous traits of these faces, we use a system that was trained to “see people like people” [168] to rate faces from the CelebA dataset on continuous traits. [9] had workers judge 40 social traits of 2,222 faces from the the MIT 10k US Adult Faces Database using Amazon Mechanical Turk (AMT). Half of the traits have positive connotations and half have negative ones. Subjects judged traits such as trustworthiness, attractiveness, responsibility, aggressiveness, etc. Human judges are highly correlated with each other on a number of these judgments such as trustworthiness and attractiveness [187]. These subjective judgments were then learned by regression on the deep layers of a pretrained object recognition network[168]. The regression network is highly correlated with human judgments for traits where the human judges are highly correlated with each other. Hence, in this work, we only use traits for which human judges are correlated with each other at a level above 0.5.

The 2,222 labeled faces from the MIT Faces Database are not enough to train a robust generative model. Hence, we used the trained regression network to judge the social traits of 190,000 faces from the CelebA dataset [114] and use the predicted trait ratings to train ModifAE. Example faces and their predicted ratings are shown in Figure 4.1.

Table 4.1. AMT verification of our collected dataset

Attribute	Chose “correct” member of the pair
Aggressive	0.9509
Emotional	0.9234
Responsible	0.7783
Trustworthy	0.8780

To verify the effectiveness of this dataset, we ran an AMT experiment checking how our predicted values align with human judgments in four attributes: aggressive, responsible, trustworthy and emotional. Humans have high agreement in these traits. For each trait, we

Figure 4.1. Examples of CelebA faces and their trait predictions.

picked 40 images rated highest by the prediction network [168], and 40 images rated lowest. Next, we form 40 pairs by picking one image from the high rating group, and one image from the low group. We then ask AMT workers which face better exemplifies the corresponding trait in the each pair, (e.g. which one looks more trustworthy in this pair). Each trait's 40 pairs are evaluated by 30 workers. Then, we calculated the overall likelihood (among all the workers and among all the pairs) that the face of higher predicted score is chosen by human subjects for each attribute. As can be seen in Table 4.1, all the attributes predicted by the regression network [168] align well with human judgments. Therefore, we consider the predicted scores as being roughly equivalent to human judgments of those faces.

Figure 4.2. Examples of ModifAE at training time (real label on left) and usage times (modified label on right).

4.3.2 Architecture

The ModifAE architecture consists of two (image and trait) encoding pathways which fuse in the middle of the network before feeding into a single image decoder. There are three pieces which combine to make the single ModifAE network: image encoder, trait encoder, and image decoder.

The image encoder and decoder are identical to the encode and decode portions of the StarGAN generator network [26]. StarGAN's generator is a convolutional autoencoder with residual blocks and a bottleneck in the middle. We use the first half of this network (including the bottleneck) as the image encoder. We use the remainder of the network as the image decoder.

The trait encoder takes a 1-dimensional set of traits, runs a single dense layer with Leaky ReLU activation, and reshapes the output to create a vector of the identical shape as the output of the image encoder.

The outputs of the trait and image encoders are then combined into a single latent representation of the image and ratings. 50% dropout is applied to the values from the image encoder, which is then averaged with the trait encoder output to arrive at the combined latent representation.

This combined latent representation is the same dimension as the image encoder output, and thus is also the same dimension as the image decoder input. So, it is passed to the image decoder to arrive at the single output image. The architecture is depicted in Figure 4.3.

While training our model, we use each image, accompanied by its known trait values inputs. The model learns a combined latent representation for the image and traits in order to autoencode the image well. However, during test time, the input trait values can be any desired value for the modification. Because the decoded image includes shared trait information, changing the traits results in a modified output image.

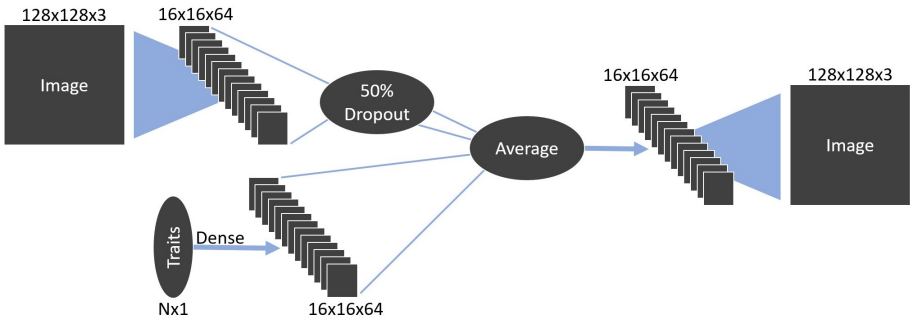


Figure 4.3. General illustration of ModifAE architecture.

4.3.3 Training

ModifAE is only trained on an autoencoding task. We train ModifAE using the Adam optimizer [92] and train for 100 epochs on CelebA images [114]. The objective is to optimize a single loss function based on two terms. We use the L_1 loss on the image autoencoder. We also optimize the L_1 loss between the trait encoder and image encoder. The total loss is:

$$L = \frac{1}{N}\sum_{p=1}^{N} |x_p - AE(x_p)| + |E(x_p) - E(y_p)| \tag{4.1}$$

where x_p is the p^{th} image example, y_p is its trait vector, $E(\cdot)$ is the result of the trait or image encoder, and $AE(\cdot)$ is the output of the full-architecture autoencoder. At training time, the objective is to take in an image and its traits, encode each independently, arrive at the same latent vector, then combine them and decode to arrive at an image identical to the input. No form of adversarial or cycle training is necessary. Despite this, the trained network can modify images without obscuring identity traits.

4.3.4 Why It Works

The image encoder compresses the image down to a bottlenecked latent space, where higher level features about the image are encoded. Simultaneously, the trait decoder projects the given traits to the same latent space, encoding an imagined face with those ratings. These equally shaped latent spaces are then combined and decoded to an output image.

However, 50% dropout is applied to the information coming from the image encoder's latent representation. Therefore, at training time, faithfully reconstructing the image is reliant on information coming from the trait encoder.

The trait encoder therefore learns to mimic average latent distributions of images with the provided ratings. This means that the information provided by the trait encoder can fill in the gaps about the image introduced by dropout in the pathway from the image itself.

So, as training converges, the network learns a joint latent space shared by the image and trait encoder, where holes in image information caused by dropout can be filled by trait information.

At test time, an image can be passed in with any desired traits. The trait encoder estimates the latent space for images with those traits, and fills in holes in the image information. When this combined information is decoded, the output image resembles the original but appears to have changes according to the provided traits.

4.4 Results

Here, we provide examples of ModifAE's performance on the novel continuous image modification task. We qualitatively compare ModifAE's and StarGAN's results for the same task, quantitatively compare the modification effectiveness with a user study, and numerically compare the ModifAE architecture with recent image to image translation methods.

4.4.1 Qualitative Results on Continuous Modifications

Multi-Trait Traversals

As emphasized by [26], the ability to modify multiple traits with a single model is important. Here, we show that ModifAE is capable of making continuous modifications on multiple traits (see Figure 4.4).

For this experiment, we trained ModifAE on two traits: "attractive" and "aggressive." The picture in the upper left corner is the original. Looking at the (0,0) point in the left set of

results (unattractive and not aggressive) the man's mouth is fairly neutral, and his features are not very pronounced. As attractiveness and aggressiveness increase, the angles of the face become sharper, there is more definition of features like eyes and eyebrows, and the smile shrinks.

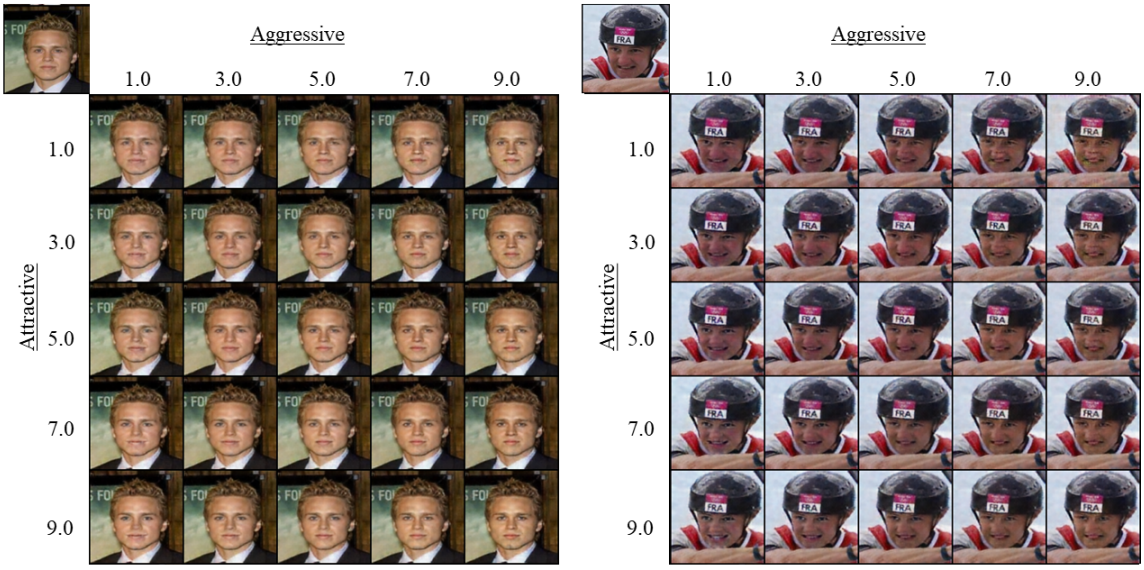


Figure 4.4. Continuous value, multi-trait image modifications by ModifAE.

Qualitative Comparison to StarGAN

To compare our model to StarGAN [26], we binarize the continuous traits by doing a median split on the continuous-valued traits and trained StarGAN on these two groups (low and high). The results are shown in Figure 4.5. StarGAN's generator architecture has a larger bottleneck and includes residual layers, so the model easily produces higher resolution images than ModifAE. However, the faithfulness of the modification of the two models is worth comparing.

Quantitative Comparison to StarGAN

To verify the quality of these continuous subjective trait modifications, we assessed human interpretations of the faces. For this, we perform an AMT experiment where we consider two traits: "trustworthy" and "aggressive." For these traits, we present participants with a

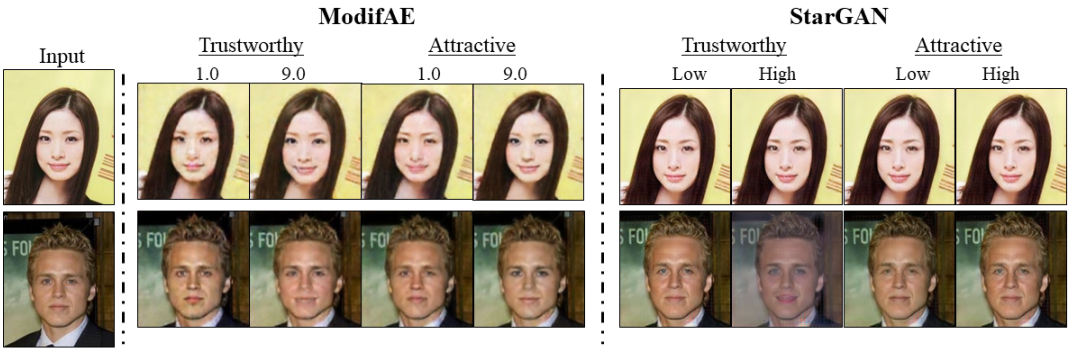


Figure 4.5. Comparison of ModifAE and StarGAN modifications.

sequence of 120 image pairs. They are asked to pick which image most exemplifies the trait.

Each sequence contains 10 ground truth pairs and 110 modified pairs. Each subject was randomly assigned to either ModifAE- or Stargan-modified images, in order to avoid judgments influenced by image resolution. A "ground truth" pair denotes a pair in which both images are the original unmodified images from the CelebA dataset, and these pairs were used to verify that workers were paying attention to the task. Since we have validated that our predicted trait scores align well with human judgment, we know the putative "correct answers" for these pairs. The overall ground truth results are between 85 and 96%, so we didn't discard any subjects.

For the modified pairs, we generated two types of pairs: (1) same-face pairs, and (2) different-face pairs. The same-face pairs are modifications of the same photo, one in which the target trait is increased, and one in which it is decreased. To make a different-face pair, photos of two different people with similar scores in the target trait are chosen and then one is modified to increase the trait, and the other is modified to decrease the trait. Both the same-face and different-face sets contain 45 pairs in which no image is repeated. However, to verify the workers' consistency, 20 of the 90 modified pairs were repeated to result in 110 modified pairs.

We calculate the fraction of subjects that chose the image with the increased trait. This is shown in Table 4.2. This can be used as a proxy for performance because higher values indicate more human agreement with the model's modifications. To verify the data collected in each model's experiment, we performed a group t-test and found no statistical differences between

Table 4.2. Human evaluation of modified images

Attribute	Performance	**ModifAE**	StarGAN
Aggressive	Ground Truth	0.94	0.95
	Same Identify (S)	0.84	0.82
	Different Identities (D)	0.66	0.59
Trustworthy	Ground Truth	0.86	0.88
	Same Identify (S)	0.80	0.39
	Different Identities (D)	0.60	0.52

the ground truth performance in the ModifAE and StarGAN experiments. This shows that the results between models are not due to an inequality in workers.

For modifications on "aggressive", ModifAE received higher scores for both same-face and different-face pairs, but the difference is not statistically significant. However, for modification on "trustworthy", there are significant differences in the overall average performance and same-face pairs performance (p¡0.001). By examining the distribution closely, we find that StarGAN is less consistent. Some StarGAN pairs receive high scores, while many other pairs cause most workers to agree on the wrong image. Meanwhile, for ModifAE, most pairs get at the majority vote for the correct image. We also observe that people are more consistent with ModifAE generated faces. On the 20 repeated pairs, 91.6% of the choices are the same for ModifAE, where for StarGAN, the number is only 45.0%.

4.4.2 Model Size

A novel aspect of ModifAE is its ability to modify images in a single forward pass, based purely upon learned weights, despite only having one network. Other models that can modify images in a single forward pass are GANs. By comparison, ModifAE, as a single neural network, requires fewer parameters and less time to train. While StarGAN takes one day to train on CelebA images [26], ModifAE takes less than 12 hours. Table 4.3 shows the number of parameters required by different models training on seven traits in the CelebA dataset. The listed values are as reported in the original papers and in the parameter comparisons of [26] [138, 218].

Table 4.3. Model size for learning seven traits

Model	Parameters
CycleGAN	736M
ICGAN	68M
StarGAN	53M
ModifAE	1M

4.5 Conclusion

In this chapter, we propose ModifAE: a novel image modification network that can edit continuous traits in a single forward pass, based solely on learned weights. ModifAE does not require training multiple networks or performing multiple steps for image modification. Instead, a single network is trained to autoencode an image and its traits through the same latent space, implicitly learning to make meaningful changes to images based on trait values. Our experiments show that ModifAE requires fewer parameters and takes less training time than existing forward-pass methods [218, 138, 26]. In addition, we computed and verified novel continuous subjective trait ratings for CelebA faces. We will make this augmentation to the CelebA dataset available upon paper acceptance. Finally, we demonstrated that ModifAE makes more meaningful continuous image traversals than equivalents generated with a state-of-the-art method [26]. Theoretically, ModifAE is not limited to face trait modification, and the resolution of output images can be greatly increased. ModifAE shows great promise as an easy-to-train model for the general task of image modification.

4.6 Acknowledgement

Chapter 4 (modifAE), in part, is a reprint of the material as it appears in Proceedings of the 41st Annual Conference of the Cognitive Science Society (COGSCI'19). (Chad Atalla, Amanda Song, Bartholomew Tam, Asmitha Rathis and Garrison Cottrell. "Modifying social dimensions of human faces with ModifAE"). The dissertation author was the co-primary investigator and

co-first author of this paper.

Chapter 5

To Dye or Not to Dye : The Effect of Hair Color on First Impressions

5.1 Introduction

First impressions play an essential role in our daily social interactions. When people encounter a stranger, they spontaneously form inferences about personality, intentions and mental states of the person [184]. As human faces contain rich social signals about emotions, intentions and attitudes, first impressions are most influenced by the person's face. Based on these first impressions, we decide whether to engage in social interaction with them, what type of social actions to perform, and what relationship to establish with them.

In the digital age, people are uploading and browsing billions of photos online. Therefore, in this study, we focus on first impressions of facial photos, and aim at illustrating how hair color affects people's perception of facial photos. Hair color is an easily modifiable physical trait, compared with other facial structure-related traits, such as eye size, or face-width-to-height ratio. It is also a factor independent of a person's identity. Therefore, the conclusions we draw can guide people to make conscious and well-informed decisions about their appearance. This study also sheds light on the potential stereotypes and bias regarding hair color and personal traits.

How hair color affects people's impression and social interactions has been examined by psychologists in field studies as well as in labs. One study examined how hair color and make-up affect evaluation on female capability and salary [98]. Females are rated as significantly

less capable with blonde or red hair as compared to brown hair. However, this study used only one female identity (changing hair color and make-up). Another study investigated the role of skin tone, hair color and hair length on the perception of female attractiveness [176], however, they used line drawings rather than real photos, and therefore lack ecological validity. Some studies have found that there is a "dumb blonde" stereotype of female faces that exists among male raters, but not female raters [199]. [64] summarized a number of studies examining the role of hair color on behavior (e.g., amount of money raised by a fundraiser, hitchhiking success rate, and tips a waitress receives). They conducted studies on courtship behavior by asking subjects to wear different wigs to a bar and comparing the courtship solicitations they received. They conclude that blond women receive more solicitations while red-headed men receive more refusals.

Our methods differ from the previous ones in the following aspects: (1) We use a GAN model (AttGAN [69]) is trained to generate a large number of realistic-looking faces varying in gender, poses, expression, hairstyle and background, and therefore have unprecedented ecological validity. (2) The advantage of a GAN model is that it can take any face image as input and generate a realistic face that is identical except for hair color. Therefore, we can easily scale it up and apply it to arbitrary faces. (3) Rather than measuring one simple behavioral metric, we select three important impression traits reflecting various aspects of the impression space: attractiveness, trustworthiness and intelligence. (4) Previous studies did not report or control the original hair color of the people who wear the wigs to whom they apply hair color changes. However, our study explicitly controls for that, allowing us to perform an in-depth evaluation of the effect of hair color changes given the subject's original hair color.

We find that observed hair color affects impressions differently, depending on the subject's original hair color. For people with dark hair (black or brown), changing into blond overall induces a decrease in attractiveness, intelligence and trustworthiness ratings. However, for people starting with blond hair, the direction of the influence of changing into a darker color could go either way, depending on the individual features of the face.

5.2 Hair color rating experiment

5.2.1 Image stimuli generation

The novelty of our image generation method is that we are able to generate images of the same identity with different hair coloring, keeping everything else as similar as possible. To achieve this, we employ a generative adversarial network model: The AttGAN model [69]. The model is trained on the CelebA-HQ dataset, which contains thousands of high-resolution face images [86]. The original dataset is annotated in hair color (three categories: black, brown and blond). We use a AttGAN model trained on Celeb-HQ and it successfully changes hair color into one of the three categories, while keeping the person's identity, facial structure and other features intact.

We visually inspect all the AttGAN-generated images and remove those with major artifacts (such as loss of hair texture, or change in the background). Apart from the hair regions (including hair, eyebrows and facial hair), AttGAN sometimes also applies changes to eye color. It seems that AttGAN picks up the natural eye-hair color correlation that exists in natural faces.

Since the factor we focus on is only hair color, we make sure the eye color remains the same across manipulations by cropping and copying the original circular iris region across generated faces of all three hair colors. To visualize the changes applied to the face images, we generate a pixel-wise difference map between the image with its original hair color and the images with hair color change. For example, in Figure5.1, the first row shows the images in the following order: the original image, AttGAN generated black, brown and blond hair images. The second row shows the pixel-wise difference between the original image and the generated images in the first row, as a color image. We observe for the blond-hair image, both the hair regions and eyebrow regions are changed. For black color, mainly the hair regions are manipulated. The last two rows present similar manipulations on a male face. In general, the major changes are constrained to hair regions (this may include facial hair in male faces, and eyebrow color changes in both genders).

We select 60 unique faces for each hair color (half female and half male faces) from Celeba-HQ, forming a unique set of 180 face samples. We assume the hair color from Celeba-HQ to be a person's original hair color. 175 of the 180 faces are Caucasian, and the other five are Asian .

For each face, we use AttGAN to generate face images for three hair colors. We note that if the target hair color is the same as the original hair color, the GAN reconstructs the image with minuscule changes. This process results in a total of 540 generated faces, balanced in gender and hair color. We include face samples from both young and old groups, based on the 'Young' binary label annotated in the CelebA-HQ dataset. We visually inspect all 540 images to ensure that the generated faces look natural and realistic to human subjects.

5.2.2 Experimental procedure

We conduct a first impression rating experiment on Amazon Mechanical Turk. We show participants one face image at a time, then ask them to rate their first impression of the face on one of the three traits: attractiveness, trustworthiness and intelligence. We choose these three impression traits because they are frequently studied by psychologists and social scientists, and are important in daily interactions [184]. The trait rating is on a scale of 1-9, with 1 indicating a low presence of the specified trait and 9 indicating a high presence. In each task, a participant is presented with 100 images and is asked to rate the faces on a single trait one by one (see Figure 5.2). Among the 100 faces, 10 are randomly repeated from the other 90 unique images (balanced in gender and hair color). We later measure a participant's self-consistency on the 10 repeated images as a sanity check. We remove those who do not have significant above-zero self-correlation (measured by Spearman's rank correlation), or who use no more than two of the nine scale values.

Each image is rated by at least 12 different participants on every trait. In total, we have collected data from 128 Amazon Mechanical Turkers (48 males, 80 females). Their age ranges from 20 to 60 years old. All participants were Caucasians to control for rater ethnicity.

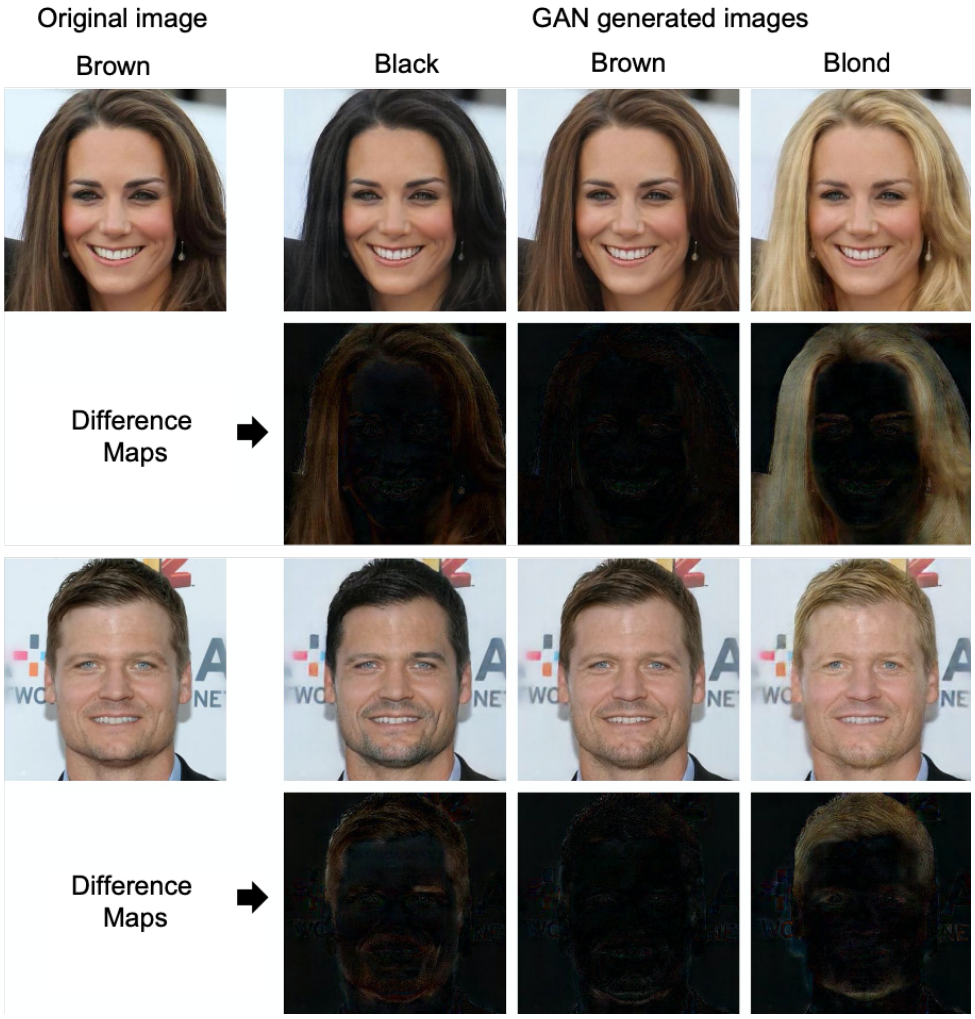


Figure 5.1. Original images(leftmost column), corresponding GAN generated images (odd numbered rows) and the difference maps (even numbered rows).

5.2.3 Results

We focus on three hair colors in this paper: black, brown and blond. Depending on the original hair color, and the color one can change into, there are six different color transforms in our study (e.g: from black to brown).

We can characterize the overall effects of observed hair color in a linear mixed effect model, treating the image as a random effect. We find that faces presented with blond hair are rated as less attractive ($b = -0.3$, $se = 0.035$), less intelligent ($b = -0.51$, $se = 0.04$), and less trustworthy ($b = -0.2$, $se = 0.043$) than faces presented with black hair. Brown hair, in turn, only

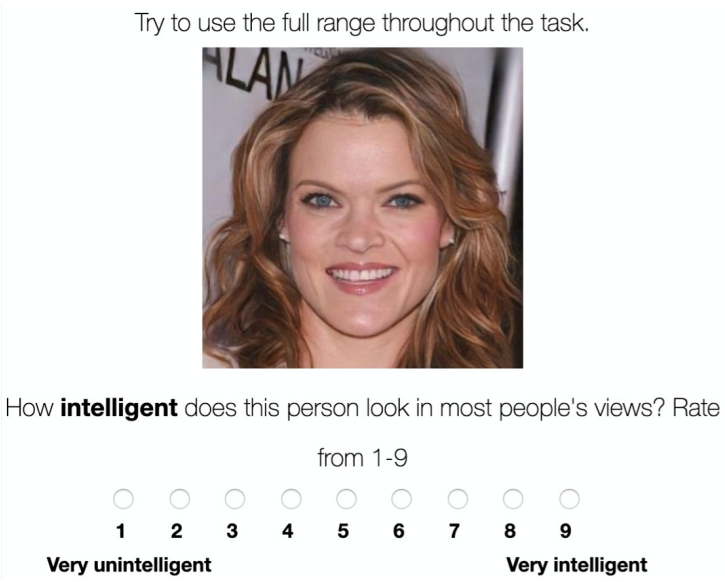


Figure 5.2. Task illustration of impression rating on intelligent.

yielded a difference compared to black hair on ratings of trustworthiness ($b = 0.2$, $se = 0.043$). Moreover, for all traits, faces presented with their original hair color received higher ratings (attractive: $b = 0.023$, $se = 0.03$; intelligent: $b = 0.23$, $se = 0.035$; and trustworthy: $b = 0.21$, $se = 0.037$). Together, these effects suggest that dying hair blond may not be a particularly felicitous strategy.

We now turn to specific transformations from one original hair color, to another candidate color. For each transformation, we compute the differences in average human ratings per trait and conduct a paired t-test. We divide the images by gender and present two heatmaps in Figure 5.3. Positive values indicate that the transformation will increase perceived value in the corresponding trait, while negative values indicate adverse effects over the population.

Changing between dark colors: As indicated from the first two rows of figure 5.3, we find that when switching between the two dark colors, black and brown, in either direction, the changes in impressions are relatively small.

Changing from dark to blond: When changing from black to blond (Figure 5.3, row three), there is a decrease in all three impression traits' ratings, especially for intelligence ratings. Similarly, when changing from brown to blond (Figure 5.3, row four), the overall trend is that it

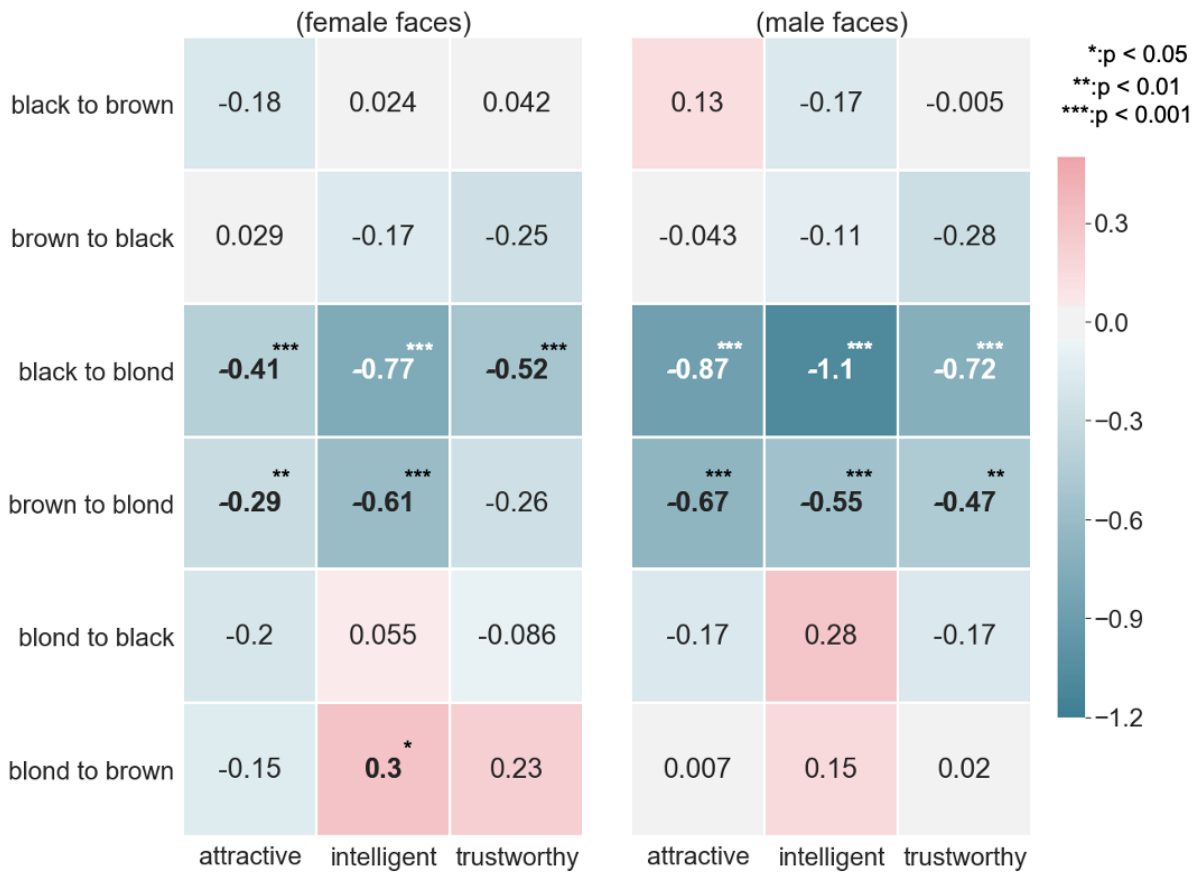


Figure 5.3. We split the images by gender, then for each trait, we calculate the differences in image ratings for each hair color transformation. Statistically significant results are marked with asterisks over the values.

also makes the person look more negative in all three traits. These results suggest that if one's original hair color is dark, changing into blond won't help improve impressions. In Figure 5.4, we present some face examples in which there are negative changes. Interestingly, we also find that for male faces, when changing from black to blond, the decrease in attractiveness ratings are significantly larger than for female faces ($p < 0.05$).

Changing from blond to dark: When changing from blond to dark colors (black or brown), the trend does not simply invert the pattern for dark-to-blond, as one might expect. For female faces, going from blond to brown can increase intelligence ratings as shown in Figure 5.3. However, overall, instead of boosting impression scores in any given trait, the effect of

transformation from blond to dark is complex, varying from person to person and differing from trait to trait. While these trends do not yield reliable differences in means across images, the changes to ratings for individual images reflect consistency across subjects, albeit inconsistency across images. The variance of rating changes across images is significantly greater than would be expected by chance for blond-to-black transitions in attractiveness $X^2 = 221$, intelligence $X^2 = 137$, and trustworthiness $X^2 = 168$; all p's< 0.001 (the same holds true for blond-to-brown transitions). This indicates that although the mean changes do not go in a particular direction, there are reliable effects for some images, although these effects go in different directions (see Figure 5.5)

Given these reliable image-level differences in the effects of hair color, we examine the changes closely in Figures 5.6, 5.7 and 5.8. Each figure focuses on one trait. In all three figures, the first two rows are examples where changing from blond to dark elicits a positive change and the last two rows are the opposite cases.

Overall, we find that when changing from a darker hair color to blond color, it significantly decreases the impression rating on all three positive traits. However, going from blond to black or brown does not always help, it only slightly increases the intelligence scores on average.

One potential confound here is that in real life, when people change hair colors, they may also change their makeup and facial hair style accordingly to fit the hair color. However, these facial changes are not generated by the current GAN model.

Another possible explanation for the overall drop in positive impression ratings might be that, due to natural selection, everyone's original hair color goes together with their facial configuration and eye color [106] as a relatively optimal choice. When switching to another hair color, it violates the potential harmony and natural co-occurrence that humans have adapted to and therefore elicits an overall drop in positive scores. Furthermore, since the faces in CelebA dataset are all celebrities, people might already be familiar with some celebrity with certain hair color and are not used to the hair color changes.

In addition, although GAN models have been successful in generating realistic faces in

general, there might be some unconscious uncanny valley created by the generated faces. Further studies are needed to evaluate the naturalness and authenticity of GAN-generated images, to rule out the possibility that the negative changes is due to deteriorated image authenticity.

5.3 A computational model of facial impressions

From the results above, we find that which hair color is better depends a lot on the specific face. Previously, a neural network model was successfully used to predict social impressions of faces [168]. Can we use a similar model to further quantify the changes associated with each specific face? Inspired by this previous work [168], we designed a neural network model that predicts social impressions of facial images to address this question.

5.3.1 Model Architecture

Our proposed network is called ImpressionNet, see Figure 5.9 for architecture illustration. We use an off-the-shelf MobileNetV2 network [158], as the backbone to extract features. All layers of the MobileNetV2 are initialized with weights from a network trained for ImageNet classification. The image features are passed to a fully connected layer with 512 nodes. The network has separate prediction sub-networks for each attribute, consisting of one hidden layer with 128 nodes and an output layer with linear activation. Each attribute's sub-network takes the 512-dimensional image features, and predicts the rating as a single continuous value. This regression network predicts a rating for each attribute and uses Mean Squared Error loss for optimization.

5.3.2 Model Training and Evaluation

We initialized the layers corresponding to MobileNetV2 in our network with weights from a MobileNetV2 trained for image classification on ImageNet. Then we fine-tune the network on our hair color rating dataset of 540 images (face images as the input, three social trait ratings as the output). We use 45% of the dataset for training, 5% for validation and 50%

for testing. The validation set is used to select the best model based on validation loss. We use a learning rate of 0.002 and train the model for 40 epochs with a batch size of 32. We refer to this trained model as ImpressionNet.

Our model's performance is evaluated by the Spearman's rank correlation coefficient between model's predicted ratings and human averaging ratings on the same set of images. As a point of reference we also report the split-half correlation between human raters, which indicates human agreement level on a certain impression trait, following similar methodology as [168].

Table 5.1. Spearman's rank correlation coefficient for proposed ImpressionNet and human raters on three social traits.

Correlation	**Attractive**	**Intelligent**	**Trustworthy**
Model with human	0.709	0.676	0.615
Human with human	0.896	0.778	0.776

Table 5.1 shows the ImpressionNet's correlation with human ratings, and human correlation with human (split-half correlation) on all three traits. ImpressionNet accurately predicts all three traits. Among the three traits, humans have the strongest agreement on attractive ratings, and the model is predicting best for this trait. This fits well with a previous paper's finding[168] that whenever humans have higher agreement levels, there are more consistent signals in the images for the model to learn.

In addition to predicting the average ratings, we also attempted to predict the degree of change in the trait ratings based on change in hair color. However, due to the limited number of hair color change pairs in our experiment, the performance was poor. This might arise for several reasons, some potentially interesting. First, the current ImpressionNet achieves an RMSE of 0.85 on the test set. This is not sufficient to predict subtle impression changes when a person's hair color is changed. Second, it may be the case that people pick up on subtle uncanny-valley effects in the GAN manipulated images, that ImpressionNet is blind to, thus missing the overall negative effects of changing hair color from the original. Collection of a larger dataset of GAN generated hair color changes and social trait ratings is required to address this. A model trained

using a larger dataset may accurately predict change in social trait ratings upon changing hair color, and can be used to inform decisions when dyeing hair at an individual level.

5.4 Discussion

In this paper, we use GAN-generated face images to examine the effects of hair color changes on impression formation. We find that of 70% of face images, all three hair colors are equally good. For the remaining 30% of the people, the effects depend on one's original hair color. For faces with original dark hair, they are more likely to undergo a negative impression change when switching to blond. For faces with original blond color, the impression change could go either positive or negative, depending on the face.

We also find that attractiveness is positively correlated with intelligence and trustworthiness for male faces, but not for female faces, which might reflect people's gender stereotypes. But since this gender difference is drawn from images of celebrities, further research is needed to test if the conclusion holds for the general population.

In this study, we did not differentiate raters' demographics. Previous research suggests that social impressions might differ depending on an individual's gender, age, cultural background and personal preference. Similarly, the ethnicity of the face may also play a role in impression perception. In future studies, we can conduct fine-grained analysis based on these factors.

The state-of-the-art GAN model equipped us with unprecedented accuracy to control the factors we want to study, and examine it under precise isolation, and variation (e.g. different hair colors). This method can be further extended to study other hair colors as well as other physical traits, such as eye color, skin tone, beard styles and glasses.

One potential confound is that the GAN-generated images have some small glitches that unconsciously elicit incongruent perception in participants, and therefore induce lower ratings on them on average. Future work is needed to further improve the GAN generated image quality and authenticity.

Our current computational model is able to accurately predict impression ratings and correlates well with human raters. However, it cannot predict the changes in impressions arising from changes to hair color. In the future, with a larger dataset with more traits, we hope to teach the model to give personalized recommendation by accurately predicting impression changes associated with different physical traits.

5.5 Acknowledgement

Chapter 5 (hairGAN), in part, has been accepted for publication of the material as it will appear in Proceedings of the 42nd Annual Conference of the Cognitive Science Society (COGS'20) as a poster. (Amanda Song, Devendra Pratap Yadav, Weifeng Hu, Garrison Cottrell and Ed Vul. "To Dye or Not to Dye : The Effect of Hair Color on First Impressions"). The dissertation author was the primary investigator and co-first author of this paper.

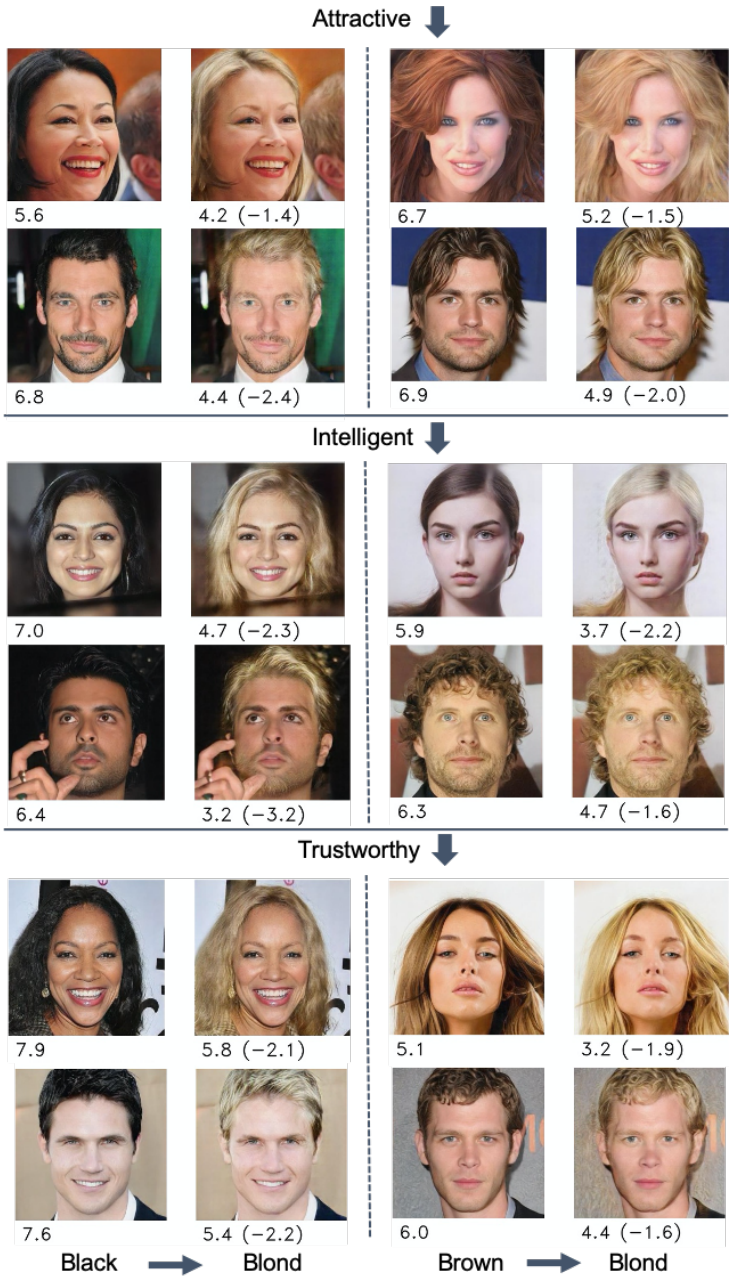


Figure 5.4. Switching from dark to blond hair color generally yields lower perceived attractiveness, intelligence, and trustworthiness. The two leftmost columns show a transition from black to blond, and the rightmost columns show a transition from brown to blond.

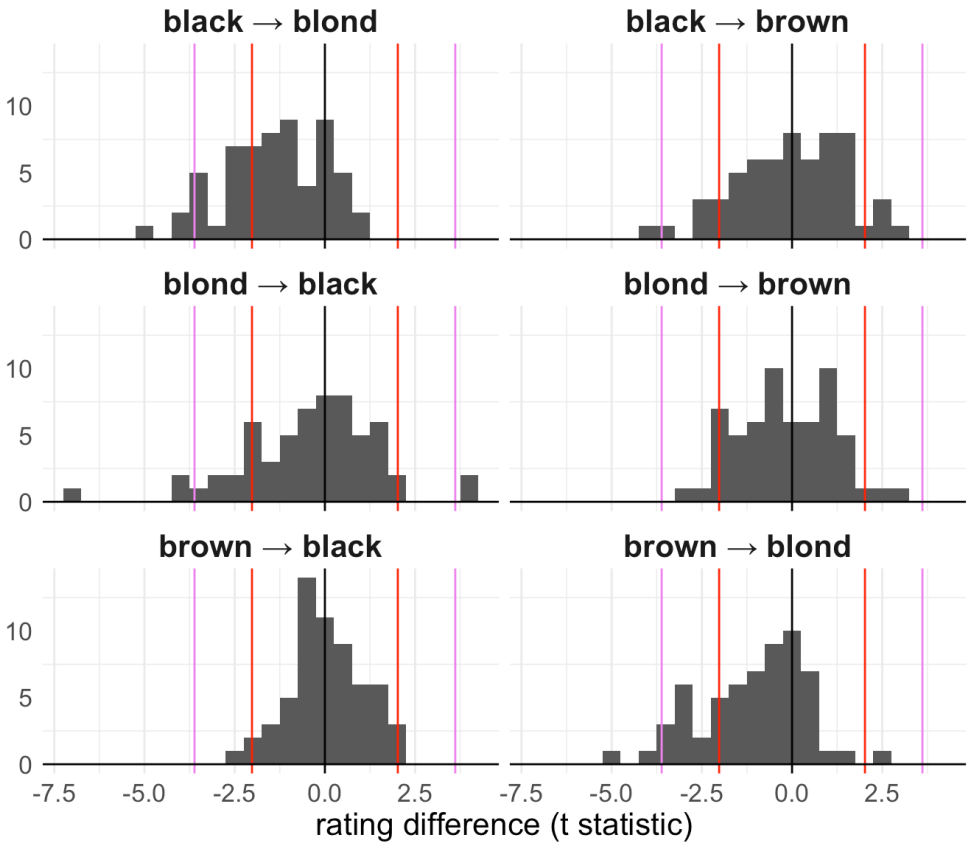


Figure 5.5. Histograms of t-statistic for image-level differences in attractive rating for individual faces, grouped into 6 hair color changes. The red line denotes 95% bounds on 0, and pink line denotes the same after Bonnferoni correction for the number of images considered.

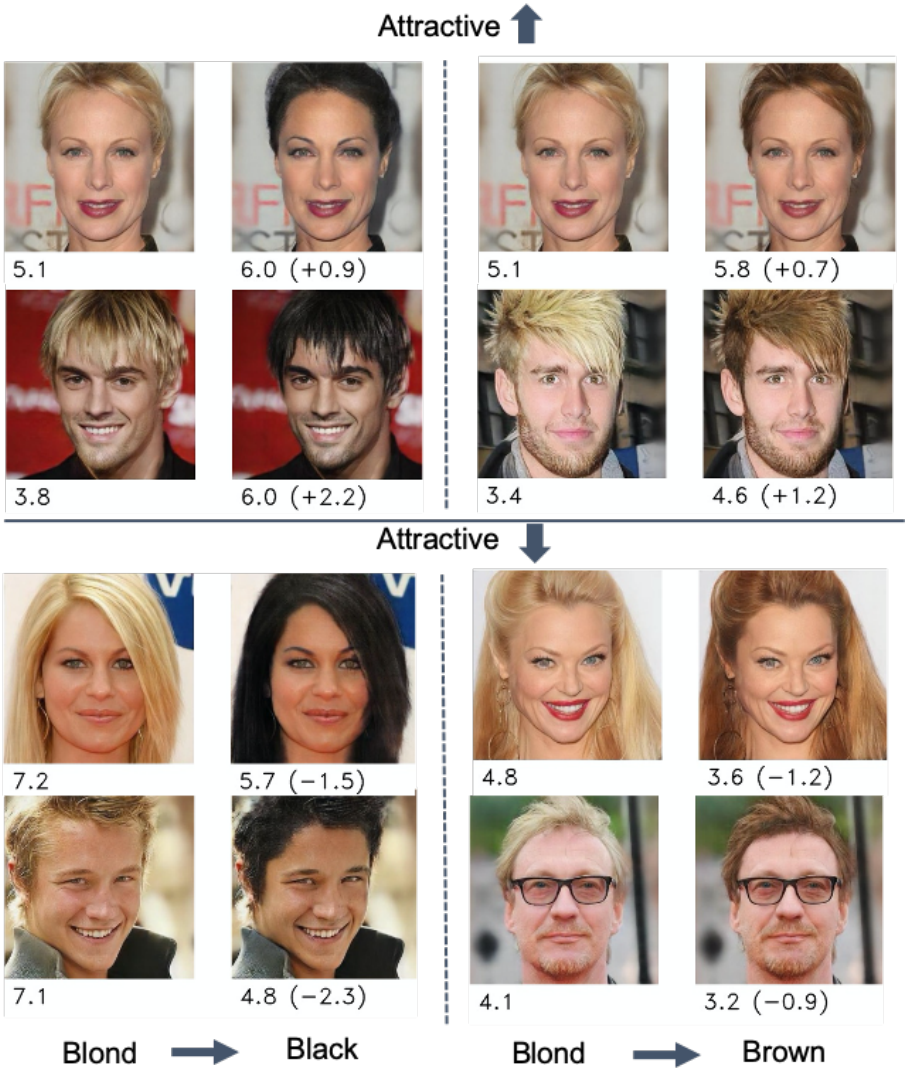


Figure 5.6. Ratings on attractiveness impression change when switching from blond to darker hair colors. The figure follows a format similar to Figure 5.4. We observe that the impression change is not unilateral and the depends on the individual.

Figure 5.7. Intelligence impression changes when switching from blond to darker hair colors. The figure follows a format similar to Figure 5.4. Depending on the individual, we may observe a large change in both directions

Figure 5.8. Trustworthiness impression changes when switching from blond to darker hair colors. The figure follows a format similar to Figure 5.4. Depending on the individual, we may observe a large change in both directions

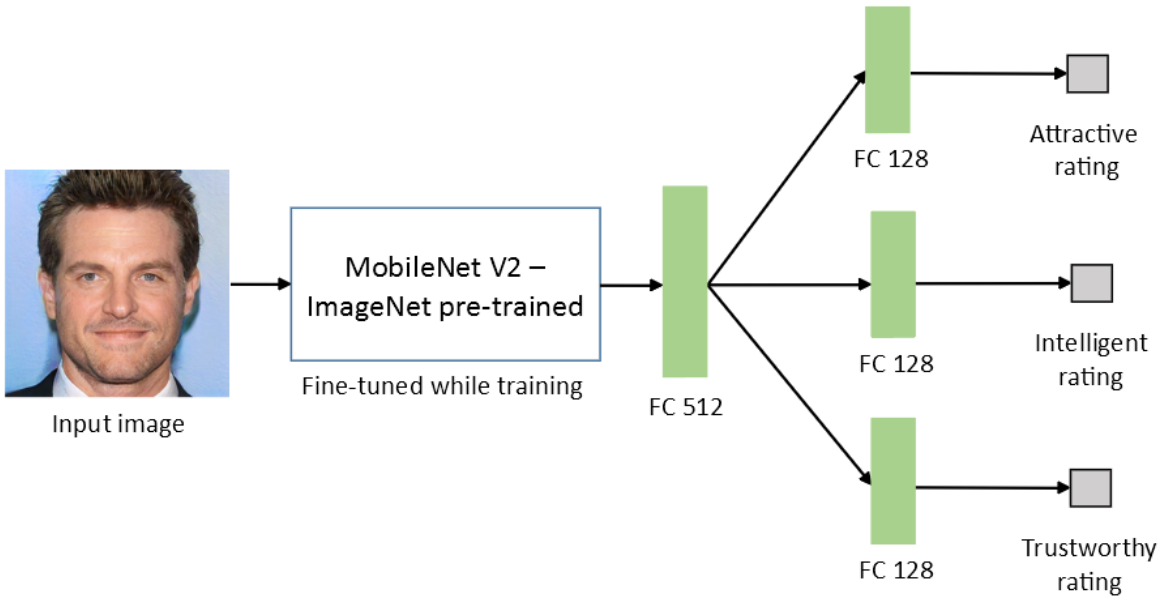


Figure 5.9. Architecture of our convolutional neural network model to predict social trait rating. FC denotes fully connected layers.

Chapter 6

Do you see what I see? A Cross-cultural Comparison of Social Impressions of Faces

6.1 Introduction

Although we are told not to judge a book by its cover, we nonetheless do it frequently when we see people for the first time. At first sight of a new person, our brain automatically forms impressions of them – how trustworthy are they? how kind? what is their social status? Even if these spontaneously formed social impressions are not objectively true [130] (consider the case of Ted Bundy!), they nevertheless affect crucial aspects of our lives, including interpersonal relationships, hiring and financial decisions [143], even legal judgments [201] and electoral outcomes [183, 184].

Regardless of their dubious accuracy, people have a fairly high agreement in the facial impressions they form [51]. This agreement is also reflected in the image-level facial features that drive impression formation, such as the apparent age, gender, race, and expressions of the face [45, 1, 210]. This agreement also arises in the correlation structure among the impressions of different traits that seem to fall along three factors: warmth, competence, and youthful-attractiveness [184, 174].

Despite these universal aspects of facial impressions, the impressions we form are also influenced by the cultural background of the viewer [184]. This should be no surprise. Research suggests that culture even shapes visual perception [127], and it certainly shapes our social

norms, expectations, and values. For instance, East Asians have been characterized as being more collective and holistic, whereas Westerners have been more individualistic and analytic [75, 133]; perhaps this would make friendlier looking people seem more capable to Asian viewers. Moreover, culture also influences our eye movements when we look at faces [16], which may mean that different facial features will be more salient to viewers from different cultures. Altogether, cultural differences in facial impressions seem entirely plausible, and their social importance may be increasingly significant, given the preponderance of face-to-face international interactions over video conferencing and social media.

Previous studies of cross-cultural facial impressions have identified similarities and differences in a number of individual traits such as attractiveness [34] and intelligence [97]. Yet, most prior studies used a small set of strictly controlled face stimuli, limiting the generalizability to everyday face photos with real-world variations. Furthermore, previous studies explored one trait at a time with different face stimuli, compromising any across-trait comparisons in cultural agreement levels. Bridging this gap requires large-scale cross-cultural studies of many impression traits using a large set of real-world facial images.

More recent work suggests some systematic differences between Chinese and British raters along three previously identified impression dimensions [174]. While this research indicates that there are substantial cross-cultural agreements in warmth and attractive dimension, and the study uses visualization to illustrate the differences in capability dimension, this approach has not attempted to quantify how facial features drive the cross-cultural differences in impression formation, therefore left the mediating mechanisms unaddressed.

Here in this study, we endeavor to understand the cross-cultural universals and idiosyncrasies of facial impressions systematically and to illustrate the missing link between mediating factors, i.e., facial features, and cross-cultural differences, i.e., how different cultural groups use the same set of facial features differently to form impressions. To do so, we compare how Chinese Asians and American Caucasians (henceforth, Asians, and Caucasians, with the country understood) form impressions of 18 traits for each of thousands of real-world Asian and

Caucasian face images. We consider 18 social impression traits that cover three major categories: (1) warmth related traits, such as warm, happy, friendly, trustworthy, extroverted, humble, calm and kind, (2) physical appearance appraised traits, such as attractive, masculine and healthy, (3) capability related traits, such as capable, diligent, high-social status, intelligent, powerful, responsible and successful.

Statistical analyses show that Caucasians and Asians disagree most on capability related traits, especially on "responsible" and "successful". The apparent age of the face is one important mediating factor behind the cultural differences: Caucasians rate older people as more successful, responsible, and humble. Apart from age, we also look at a broader list of relatively objective facial features, such as the gender, ethnicity, and smiling level of a face, and establish a linkage between in total nine facial features and impression traits. By fitting Lasso regression models, we capture the interesting culturally universal and specific trends, at the broad impression dimension level, as well to each particular impression trait level.

Our study provides insights into the mediating factors of cross-cultural perception of faces and suggests directions to future researchers of the critical high-level facial features related to first impression formation. It demonstrates different potential stereotypes and bias towards certain facial features, whose cultural roots are open for future work to investigate.

6.2 Methods

In this study, we aim to compare Caucasian and Asian cultural differences in the social impression perception of American Caucasian and Chinese Asian faces. To this end, we had Caucasian and Asian participants rate their first impressions of thousands of Caucasian and Asian faces on 18 socially relevant traits.

6.2.1 Face Images

We selected 1,836 Caucasian faces from the US 10K Adult Database [8]. For Asian faces, we followed a procedure similar to [8] and collected Asian faces from the internet. We

gathered the most frequently used Chinese first names and last names for both genders, and then used the combination of first and last names (in Chinese characters) as the keywords to search images from Microsoft Bing Image search engine. We then downloaded the first few face images that were associated with the name combination. After the original images were downloaded, we cropped the face regions out of the images with a face detection library [91]. Following the same filtering procedure as described in [8], we discarded any images that either: (1) with a face region resolution smaller than 200 px $\times$ 200 px; (2) depicted celebrities (to the best of our knowledge); (3) where at least more than half of the face was occluded; or (4) depicted an infant. After preprocessing, we kept 1,738 Asian faces. Figure 6.1 shows a few examples of the Caucasian and Asian face stimuli.

Figure 6.1. Examples of Caucasian and Asian face stimuli.

Annotation of high-level facial features

Previous research has shown that multiple facial features, such as age, gender, ethnicity, and expression of the face, are essential for facial impressions [45, 1, 210]. To quantitatively examine the influence of facial features on impressions, first, we need to annotate high-level facial descriptors of all the faces in our dataset.

Since these annotations are not readily available, our strategy is to train classifiers on a large scale dataset that comes with these labels, then applied to trained classifiers in our own dataset. We utilize another large scale dataset, CelebA-HQ[87], which contains 30,000 high-quality face images labeled on 40 high-level features. We train a Convolutional Neural Network (CNN) classifier based on [216] by fine-tuning a ResNet-50[68] model pre-trained on ImageNet.

Based on the classifier's performance, we remove facial features that had poor classification accuracy. We also removed the highly overlapping facial features. After the pre-filtering, we kept the following eight high level binary facial descriptors: *gender, ethnicity, smiling, bushy eyebrows, high cheekbones, wearing lipstick, wearing eyeglasses, having beard*. The average classification accuracy for these eight traits is 82.3%. We use this classification model to predict facial features for all the faces in our dataset. Additionally, we also obtain an estimate of the face's age using Amazon Recognition API. In addition to the eight binary facial features, we have this additional continuous estimated-age variable to form a nine-dimensional feature vector for every face image in our dataset.

6.2.2 Social Impression Traits

We used 18 social impression traits that align with the three key dimensions commonly found in prior research on first impressions of faces [174, 184]. (1) warmth and approachability related traits: friendly, happy, kind, trustworthy, extroverted, humble, and warm. (2) attractive/youthful and physical appearance-based traits: attractive, masculine, and healthy. (3) competence related ones: calm, capable, diligent, (of) high social status, intelligent, powerful, responsible, and successful.

6.2.3 Impression rating collection

In the online experiment, we ask participants to indicate their first impression of an image on a specific trait by providing a rating on a scale of 1-9, as shown in Figure 6.2. We asked people what they think others would rate the face to solicit their own preferences, as suggested by our unpublished data to have an effect on reducing social desirability biases when offering potentially controversial opinions. Participants saw multiple faces in a sequence and rated one face at a time. For Asian participants, the task instructions and all impression traits were translated into simplified Chinese and then back-translated into English to ensure that the meanings are maximally preserved.

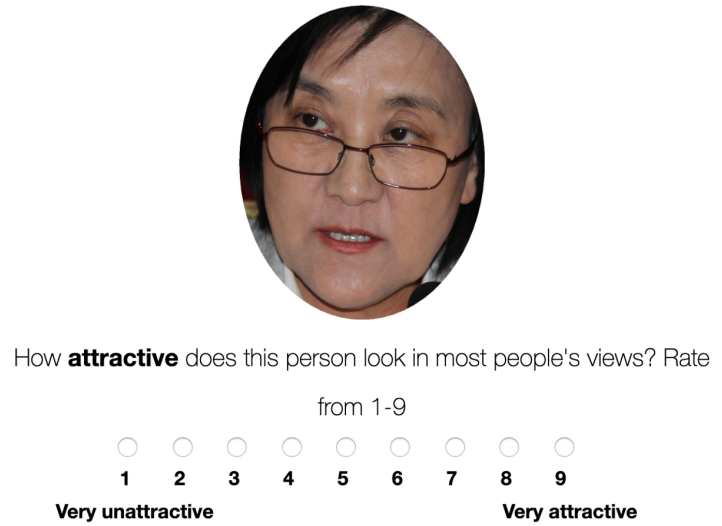


Figure 6.2. First impression rating task page.

Since rating social traits is a subjective task, we designed a screening mechanism to ensure participants were paying attention to the task. The screening consisted of 20 randomly selected unique faces of the same ethnicity and a randomly-selected social trait to rate. The 20 faces were presented, then they were shuffled and shown again, resulting in a 40-trial sequence. If a participant's reliability was significantly above zero (as measured by Spearman rank correlation of test/retest ratings), and they used at least three different scores from the 9 point scale, the participant passed the "reliability test". Otherwise, the participant's data is not used. For each rater ethnicity group, we collected at least ten ratings per image-trait combination.

6.2.4 Rater recruitment specifics

Due to constraints from the different data collection platforms in the two countries, there are differences in the specific collection protocols, as described below.

US raters were recruited from Amazon Mechanical Turk [109] (they self-reported as Caucasians and live in the US). Once a Caucasian rater passed the initial screening phase, he or she can revisit the rating task multiple times. Each time, they were randomly assigned to rate a specific trait on a different set of faces. In each task, there are 100 faces of the same ethnicity, among which 90 are unique, and the other ten faces are randomly drawn from the 90 unique

ones. We found the reliability was adequate for subjects that passed the first screening, so we did not analyze the ten repeated faces further.

Since there is no exact equivalent version of the Amazon Mechanical Turk platform in China, we recruited Chinese participants via the data100 website, (the web page address is https://www.data100.com.cn) as well as online volunteer sourcing. The constraints on the data collection platform and volunteering mechanism meant that each participant accessed the task only once. Therefore, there is no chance to screen them first, then re-invite them for multiple tasks. So we merge the screening process into the 100-face-trial rating task, such that the first 40 faces are for screening purposes, and the remaining 60 faces are unique and different from the first 40. If a Chinese rater passes the screening phase, he or she will rate all the 100 faces; otherwise, the task will terminate after the first 40 screening trials.

In total, we recruited 428 (254 females) Caucasian raters (the median age range is 30-39 years old). Due to the data collection platform differences, many more Chinese Asian participants were recruited: 23,304 in total, 14,338 of them are females, and the median age range is 20-29 years old.

6.3 Dataset Analyses and Results

6.3.1 Group Mean Analyses

How do Asian and Caucasian participants rate faces differently on average? Which ethnicity group would give higher ratings to faces on which impression traits? Is there an own-group advantage, i.e., people will give faces of their own ethnicity higher evaluations in the case of facial impression judgments? To find out answers to these questions, we calculated Caucasian and Asian raters' average ratings on 18 impression traits. For each impression trait, we split the faces into four groups by the ethnicity and gender of the face. The results are shown in Figure 6.3. A follow-up ANOVA in Figure 6.4 further illustrates the variance explained by every single factor and the interactions among them.

We observe that Asian raters give overall lower ratings than Caucasian raters. All of the ratings in Figure 6.3, including happy, are significantly higher for Caucasians over Asians ($p < 0.01$). This trend aligns with prior results, arguing that compared to Chinese participants, European Americans tend to emphasize the positive and downplay the negative [164].

Second, we find that on average, images of Caucasians are rated higher than images of Asians across all traits ($\beta = 0.22$, se $= 0.008$), in particular for **warmth related traits** ($\beta = 0.41$, se $= 0.014$). However, smiling seemed more common among the Caucasian faces than Asian faces in our pseudo-randomly sampled image set. To correct this, we tagged whether a facial image is smiling using AWS Rekognition. We found that 75% of Caucasian images were smiling, while only 31% of Asian images were. Correcting for the effect of smiling reverses the image ethnicity effect, such that warmth related traits are rated lower for Caucasian smiling images than Asian smiling images ($\beta = -0.14$, se $= 0.017$), and lower for Caucasian non-smiling images than Asian non-smiling images ($\beta = -0.56$, se $= 0.019$). Table 6.1 shows the dramatic disparities in smiling rates and the reversal of the Caucasian advantage when smiling is controlled. This pattern of results is suggestive of raters implicitly correcting for the different base rates of smiles among Asian and Caucasian faces; thus making a smile more diagnostic for Asian faces, and a lack of smile more diagnostic for Caucasian faces. Regardless of the specific reason, the direction and magnitude of the mean difference in ratings for Caucasian images appears to be driven entirely by the preponderance of smiles in Caucasian images, not due to differences in how Asians and Caucasians are perceived.

Table 6.1. Average ratings across all warmth related traits when separating images by ethnicity and whether they are smiling.

		Asian Raters	**Caucasian Raters**	**%**
Non-smiling	Asian	4.56	4.34	69%
Non-smiling	Caucasian	4.13	3.65	29%
Smiling	Asian	5.52	6.72	31%
Smiling	Caucasian	5.60	6.35	71%

Besides warmth related traits, we can see interesting cross-cultural similarities and differences, as well as interaction patterns in the following traits by examining Figure 6.3 and 6.4 closely. For each effect in each trait, we report the Tukey HSD/Range corrected 95% confidence interval on the relevant pairwise difference.

Physical Appearance Related Traits

Attractive: Images of Caucasian females are rated as less attractive than those of Asian females by both Caucasian raters $[-0.58, -0.38]$, and Asian raters $[-0.25, -0.08]$.

Capability Related Traits

High social status: Caucasians rate males as lower in social status than females; this holds true for both Caucasian images $[-0.43, -0.23]$, and Asian images $[-0.43, -0.27]$. In contrast, Asians rate Asian males as higher in social status in comparison with Asian females $[0.02, 0.17]$, with no significant gender difference for Caucasian images $[-0.19, 0.03]$).

Powerful: Both Asian and Caucasian raters rate males of the *other* ethnicity as more powerful than males of their own ethnicity (i.e., Asians rate Caucasian males as more powerful than Asian males $[0.04, 0.26]$; Caucasians rate Asian males as more powerful than Caucasian males $[0.14, 0.34]$).

Successful: Asian raters give the lowest ratings to Asian male images (lower than images of Asian females $[-0.728, -0.5792]$, Caucasian males $[-0.6852, -0.4755]$, and Caucasian females $[-0.9, -0.75]$). No such effect appears for Caucasian raters.

Responsible: Both Asians and Caucasians rate male images of their own ethnicity to be the least responsible. Specifically, Caucasians rate images of Caucasian males as less responsible than images of Caucasian females $[-1.11, -0.88]$, Asian males $[-0.48, -0.26]$, and Asian females $[-0.35, -0.12]$, while Asians rate Asian males as less responsible than Asian females $[-0.61, -0.45]$, Caucasian females $[-0.76, -0.59]$, and Caucasian males $[-0.36, -0.13]$.

6.3.2 Consistency analysis

Impressions of faces are subjective in nature and everyone is their own judge and may hold a unique opinion. Yet, one property that makes facial impression interesting is the consensus people share to a certain degree. In other words, universality and idiosyncrasies are two sides of the same coin for first impressions. Universality, the agreement people share on a particular impression, may reflect the influence of our shared evolutionary history on our perception of impressions; idiosyncrasies, the disagreement people bear on a specific impression, on the other hand, may reflect the influence of our unique environment, including culture, personal history, and our bias on our impression formation.

For each of the 18 impression traits at hand, we can measure the consensus degree at three levels: (1) individual level: how consistent and reliable an individual is when evaluating a specific impression trait; (2) intra-ethnicity level: how much agreement people from the same ethnic group have (3) inter-ethnicity level: how much agreement Asian raters have with Caucasian raters. These three levels serve as an anchoring point to each other and offer us a vibrant glance of the convergence and divergence of opinions of various impression traits.

Individual Consistency

Since giving impression ratings is a highly subjective task, we conducted a sanity check to ensure the quality of our data. In particular, we computed the test/retest Spearman correlation [219] on the repeated trials during the screening phase for each rater. Given we have 20 repeated trials, the threshold for significantly above zero ($p < 0.05$) is 0.38. The actual average Spearman correlation is above 0.7 for both rater ethnicities, giving reassurance that our participants were self-consistent and reliable.

Intra-group Consistency

After confirming the individual level consistency, we next examine the agreement level within each rater ethnicity for each trait. We further split each ethnicity group by gender to find

out whether there is any potential gender difference in agreement levels.

We used a one-way intraclass correlation coefficient (ICC) to measure the agreement level within a group by evaluating the ratio of the variance of item random effects to the overall rating variance. Figure 6.5 shows the ICCs of each trait for each demographic participant group ranked by the overall average ICC among all impression traits. Asian raters have a lower ICC than Caucasian raters; this lower group-level consistency among Asian raters may reflect more diverse opinions about how to evaluate these social traits.

Within the same ethnicity group, there were no statistically significant differences between male and female participants. Similar to previous research [70], we found that there is more agreement for traits representing warmth and appearance-based appraisals (e.g., happy, warm, friendly, kind, attractive), than for competence-related traits (such as diligent, capable, intelligent, and powerful). This effect should not be too surprising as attractiveness, youth, and propensity to smile are much more evident in a picture than traits like diligence.

6.3.3 Inter-group Correlation Analysis

Now we address the third level of consistency, the inter-ethnicity-group consistency, to understand what impressions are perceived more universally and what impressions are perceived differently across cultures. To quantify the agreement level Caucasians and Asians have on each impression trait, we use Spearman correlation as a metric and compare the inter-group correlation levels among 18 impression traits. For each impression trait, we further ask whether the agreement level differs depending on the face ethnicity, to check if there is an in-group effect.

To qualitatively examine the differences in the ratings on responsible, successful, and humble, we rank facial images based on how differently they are rated on average by Caucasian and Asian raters, then divided by the sum of standard deviation within each rater ethnicity group. i.e., $\frac{\mu_{Asian}-\mu_{Caucasian}}{\sqrt{\sigma^2_{Asian}+\sigma^2_{Caucasian}}}$. In this way, we can see the faces that are rated higher by Asians than by Caucasians, and vice versa on several traits, as shown in Figure 6.7. For each trait, the nine images on the left are rated higher by Caucasian participants, and the nine images on the right

are rated higher by Asian participants, after correction on variance. For face identity protection, we morphed several faces together to form each exemplar we show.

6.3.4 Age-based analysis

Based on visually inspecting the extreme face examples, we see that for all the three traits, faces rated higher by Caucasians seem to be older than those rated higher by Asians. We hypothesize that age might play a role in Caucasian and Asian raters' different facial impressions. To test if this hypothesis holds beyond the few examples we see, we label the perceived age of all the faces in our database, using an age-tagging API by Amazon Rekognition. Each face is given an age estimation range, and we take the average of the range as the approximate age of the face.

We divide the face images into different age ranges (in five-year intervals), from below 20, all the way to above 55, and then we split the raters and faces by ethnicity and plot the average rating as a function of age for each rater ethnicity-image ethnicity combination for the three impression traits we examined above. As we can see from Figure 6.8, the overall trends are Caucasian raters give higher ratings to more senior people, whereas Asian raters give no higher, if not lower ratings to more senior people. And the trend holds for all three traits: successful, humble, and responsible, although with varying degrees of difference.

To further quantify the degree of differences in Caucasian vs. Asian raters' responses to age, we fit a linear regression model for Asian raters and Caucasian raters respectively for each social trait, using the age of images to predict the average rating of images. We divided the data by the ethnicity of the raters, and we split images into one of the nine categories based on estimated age from below 20 to above 55. For each rater ethnicity group, we took the average of the z-scored ratings for each rating as the final ratings. A linear regression model is then used to fit the final ratings vs. the estimated image ages (continuous). Lastly, we visualize the coefficients of the linear model using a heatmap. We plot the slope of each model on each trait for Caucasian and Asian raters in Figure 6.9. The first column shows the overall common effect, the second and the third columns show the deviation from the effect for the different rater ethnicity.

As we can see from the figure, the overall effect is larger than the ethnicity-specific effects, but there are considerable ethnicity-specific variations, most notably in attractive, responsible, successful and humble.

From this figure, we can see that for responsible, humble, and successful, Asian participants and Caucasian participants have coefficients of opposite signs, confirming our hypothesis that age is judged differently when forming impressions.

It is worth noting that our Caucasian and Asian participants have a slightly different median age range. We split participants into different age groups, to rule out the possibility that the differences we observe across cultures are due to age differences of the participants. See Figure 6.12 in supplementary material for more details.

6.3.5 Lasso regression model on social impression

Lasso regression model

To quantitatively understand to what extent these high-level facial features contribute to the perception of social impressions, we train a Lasso regression model that uses the high-level facial features to explain and predict the social impression traits.

One of the benefits of using a Lasso model is that its L1 regularization will drive coefficients of unrelated facial features to zero, therefore giving us a concise high-level picture with the remaining non-zero coefficients.

The Lasso models are trained on our cross-cultural dataset of Caucasian and Asian images, with each image rated on 18 social impressions and nine high-level facial features.

We train a separate model for social impression ratings from Caucasian raters and Asian raters, respectively. With the two trained models, we can visualize the learned coefficients to identify the most relevant facial features for every social impression trait and compare them across rater ethnicities.

Results and discussion

We evaluate our Lasso regression models using the coefficient of determination (R^2) and the Spearman's rank correlation, which is calculated between the predicted and ground truth human social trait ratings.

For Caucasian raters, the Lasso model achieves $R^2 = 0.41$ and Spearman's correlation = 0.62. For Asian raters, the Lasso model achieves $R^2 = 0.25$ and Spearman's correlation = 0.46. We observe that the higher noise in ratings from Asian subjects causes a drop in performance. We visualize the coefficients from the Lasso model trained on Caucasian and Asian raters in Figure 6.10 and Figure 6.11, respectively. In both figures, we sort the nine facial features from left to right based on their average absolute magnitude. So the more important features are on the left. Age and smiling are the two most important factors, whereas features like having bushy eyebrows or beard are relatively less important. On the 18 social impression traits, we sort them based on the four broad categories: warmth-related ones, capability-related ones, attractive-youth, and masculine. We find a similar categorization of social traits upon performing K-means clustering on the lasso model's coefficients, with K=4 using Scikit-Learn [136].

First, we look at each facial feature's overall effects on each broad category, e.g., whether age has a positive, negative, or neutral impact on warmth-related perception. We compute the average coefficient of one specific facial feature on one broad category from one rater ethnicity model. We define the overall effect positive when it's above 0.1, negative when it's below -0.1, neutral when it's in between. For most of the facial features, the effect is the same for Caucasian raters and Asian raters; there are only two exceptions. The general trends and exceptions are summarized in Table 6.2. Smiling, wearing lipstick are positive factors for warm, capability, and attractive-youth related perceptions. Apart from that, for warmth-related impressions, the other positive factors are the face "is Asian", and having high cheekbones. The negative factors are apparent age of the face and the face "is-male". For capability-related impressions, age, is male, and wearing eyeglasses are positive, and having a beard is negative. One cultural difference is that

a face is Asian hurts capability-related impressions, but only for Asian observers. For attractive-youth, bushy eyebrows are a positive factor, and age plays a negative role. For masculine perception, not surprisingly, is male is positive and wearing lipstick is negative. Age is a positive factor, but only for Caucasian people.

After examining facial features' overall effects on broad impression categories, we then zoom in to identify the culturally different patterns at a single impression trait level. For each facial feature - impression trait combination, we have a pair of coefficients, one from the Caucasian people's model and one from the Asian people's model. We compare the absolute difference of the two coefficients: if it's above 0.55 (95% percentile of all the coefficients), then we consider this facial feature has a culturally different impact on this impression trait. Based on our threshold criteria, we find Caucasians and Asians form the following impression traits differently: Being senior and wearing eyeglasses make one look more responsible and successful in Caucasian people's eyes. More senior people look humbler and more diligent for Caucasians, but of higher social status and less successful for Asian people. Age decreases attractiveness perception more for Caucasians than for Asian observers. Smiling increases the trustworthy impression more for Caucasians than for Asian observers.

Table 6.2. Facial features' overall effects on broad impression categories

Category	Positive factors	Negative factors
Warmth	Smiling, wearing lipstick is Asian, high cheekbones	Age, is male
Capability	Smiling, wearing lipstick Age, is male, eyeglasses	Beard
Attractive-youth	Smiling, wearing lipstick bushy eyebrows	Age
Masculine	Is male	Wearing lipstick

One issue that is worth noting is that since our facial features are correlated (e.g., females are more likely to wear lipsticks), due to the co-linearity, the multivariate analysis result is subject to change when we include more or fewer features. However, the general trend, as confirmed by

univariate analysis, in large support the main trends we find with Lasso multivariate regression analysis. For instance, we conducted a univariate regression on age, and although the coefficients themselves have different values, the overall effect of age on warmth related impressions are positive. In contrast, the overall effect of age on capability related impression traits are negative. The culturally unique patterns, such as Asians rate senior people as of higher social status, and that Caucasian people rate senior people as more diligent, humble, responsible, and successful also persist in univariate analysis, see Figure 6.9 for details. Apart from age, the other eight facial features' univariate analysis results are listed in the supplementary section, Figures 6.13, 6.14, 6.15,6.16, 6.17, 6.18, 6.19, 6.20.

6.4 General Discussion:

In this paper, we conduct a large-scale cross-cultural study of facial impressions, investigate the mediating factors underlying impression formations, and the cultural universals and idiosyncrasies regarding how Caucasians and Asians use these facial cues to form impressions of Caucasian and Asian faces.

First, we find that there is a significant difference between Caucasian and Asian participants regarding their group agreement levels, which suggests that Caucasian participants tend to judge most traits similarly, whereas for Asian participants, there are diverse opinions on most of the impression traits in facial images. Although we have already control the sample size with the ICC measure, the difference may or may not have a cultural root since Asian and Caucasian raters are recruited via a slightly different procedure due to practical difficulty in matching the recruitment methods strictly.

Overall, we find that Caucasians give higher ratings than Asian participants on almost all the positive traits. Caucasian faces, in general, receive higher ratings on warmth related impression traits, likely due to the fact that there are more smiling faces in the Caucasian face sample pool. Once conditioned on smiling or not, Asian faces receive higher ratings than

Caucasian ones. Among the 18 impression traits, people disagree more on competence related ones and agree more on warmth related ones. Among all impression traits, people disagree most on responsible, humble, and successful, and they disagree more on Asian faces than Caucasian faces.

By visualizing the faces that are rated most differently by Caucasians and Asian in these three traits, we spot a pattern that faces rated higher by Caucasians are older than faces rated higher by Asians in these three traits. To test if this trend holds across the whole dataset, we give an age label to every face in our dataset using an API from Amazon. We plot the average ratings on faces as a function of age for Caucasian raters as well as for Asian raters and find that the trend we observe in the few face examples indeed holds for the whole dataset. A regression model further gives a quantitative measure of the slope and validates that Asians and Caucasians respond to age differently when forming impressions of responsible, humble, and successful.

We extend our attention to more high-level facial features and probe the relationship between more facial features and social impressions. We take advantage of a fully labeled dataset to train classifiers that automatically label our own face stimuli with the high-level features of interests. We focus our attention on nine facial features: age, gender, the ethnicity of the face, smiling or not, bushy eyebrows, high cheekbones, wearing lipstick, wearing eyeglasses, and having a beard. We train two sets of Lasso models on Asian and Caucasian people's rating data, respectively, and by comparing the coefficients from the two races' models, identify the cultural similarity and difference. Regarding each facial feature's influence on the broad impression categories, the impact is the same for both cultures in most cases; the only two exceptions are: Asian faces have a counter-productive effect on capability-related impressions only for Asian raters; age has a positive influence on masculine perception only for Caucasian raters. The following results are culturally universal: smiling and wearing lipstick have positive effects on warmth, capability, and attractive-youth related perception. The detailed quantitative contribution of each facial feature on each impression can be inferred from the coefficient table. We find out the following culturally different patterns: age is used very differently by observers from two

cultures: more senior people look humbler and more diligent for Caucasians, more of higher social status yet less successful in Asian observers' eyes. Age also decreases attractiveness perception more for Caucasians than for Asians. Caucasian observers value smiling more in trustworthy impression. Lastly, wearing eyeglasses and being senior makes one look more successful and responsible, but only in Caucasian people's eyes.

It is worth further investigation to understand the deeper cultural root behind the different ways the two races use these facial features to form impressions. Previous research suggests that Asian culture has respect for the elderly, how does this cultural tradition connect with the exact impressions Asians have for older people?

As we mentioned earlier, due to practical constraints, Asian and Caucasian raters are recruited from a different channel, and their demographics could be better matched in future studies.

Our large-scale cross-cultural dataset also enables computational social scientists to build a more representative, diverse, and inclusive algorithm which can predict and modify impression based on Caucasian and Asian people's preferences. Further studies combined with GAN models can establish a more explicit causal relationship between facial features and social impression traits, e.g., adding beard directly on images to validate if beard indeed decreases capability impressions, as found in our current study.

Our dataset and statistical analyses provide new perspectives for cross-cultural studies of facial impressions. They highlight interesting patterns on how Caucasian and Asian participants use high-level facial features similarly and differently to form impressions. It advances our understanding of the mediating mechanisms underlying social impressions across cultural and provides insights to look for deeper cultural roots to explain people's impression formation patterns.

6.5 Supplementary material

6.6 Acknowledgement

Chapter 6 (cultural model), in part, has been accepted for publication of the material as it will appear in Proceedings of the 42nd Annual Conference of the Cognitive Science Society (COGSCI'20)., (Amanda Song, Weifeng Hu, Devendra Pratap Yadav, Fangfang Wen, Bin Zuo, Garrison Cottrell and Ed Vul. "Do you see what I see? A cross-cultural comparison of social impressions of faces"). The dissertation author was the primary investigator and co-first author of this paper. Chapter 6, in part is also currently being prepared for submission for publication of the material. (Amanda Song, Weifeng Hu, Devendra Pratap Yadav, Fangfang Wen, Bin Zuo, Garrison Cottrell and Ed Vul.) "Mediating mechanisms of facial impressions: cultural universals and idiosyncrasies." The manuscript's title may be subject to change. The dissertation author was the primary investigator and author of this material.

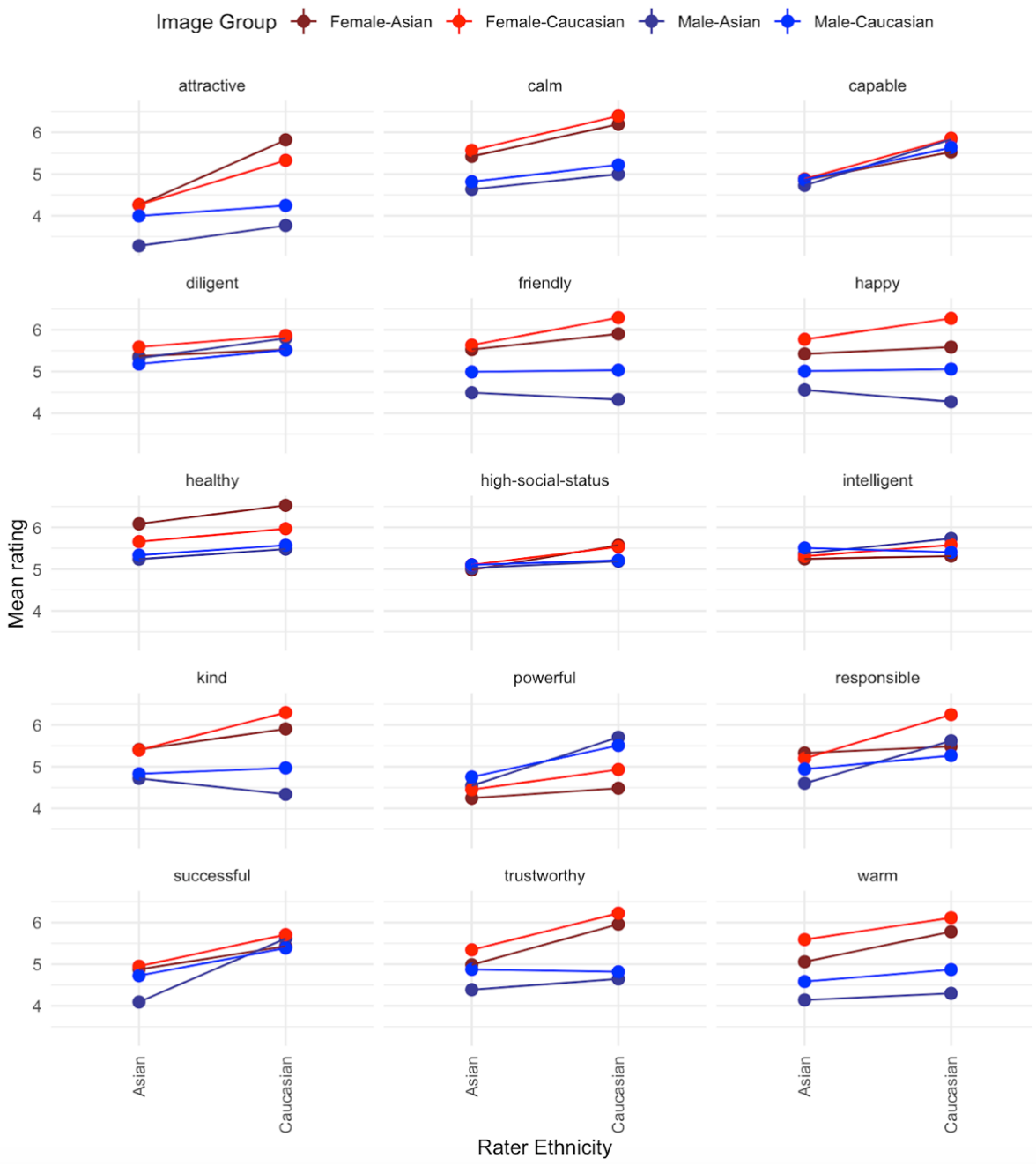


Figure 6.3. For each trait, we split the images by gender and ethnicity, and assessed Caucasian and Asians raters' mean ratings and standard errors for the 4 image groups. Overall, Caucasians give higher mean ratings and Caucasian faces in general receive higher ratings.

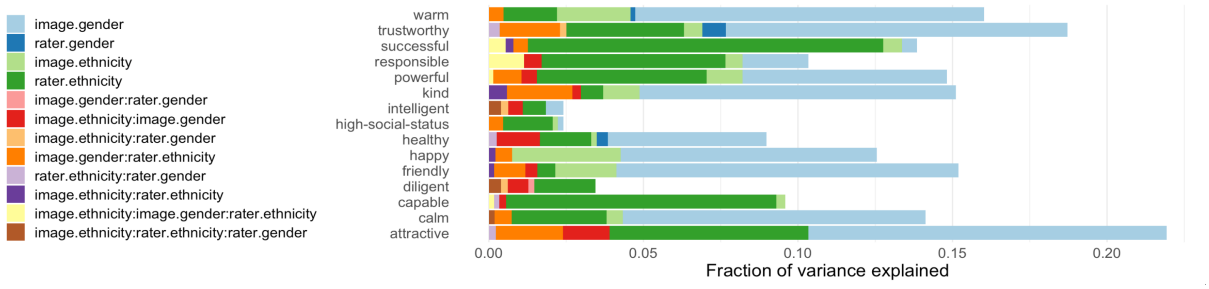


Figure 6.4. ANOVA analysis. For most features, the dominant explanatory factors are image gender (light blue; reflecting that females are rated as more attractive, warm, and friendly), and rater ethnicity (dark green, reflecting that Asians tend to give less positive ratings overall).

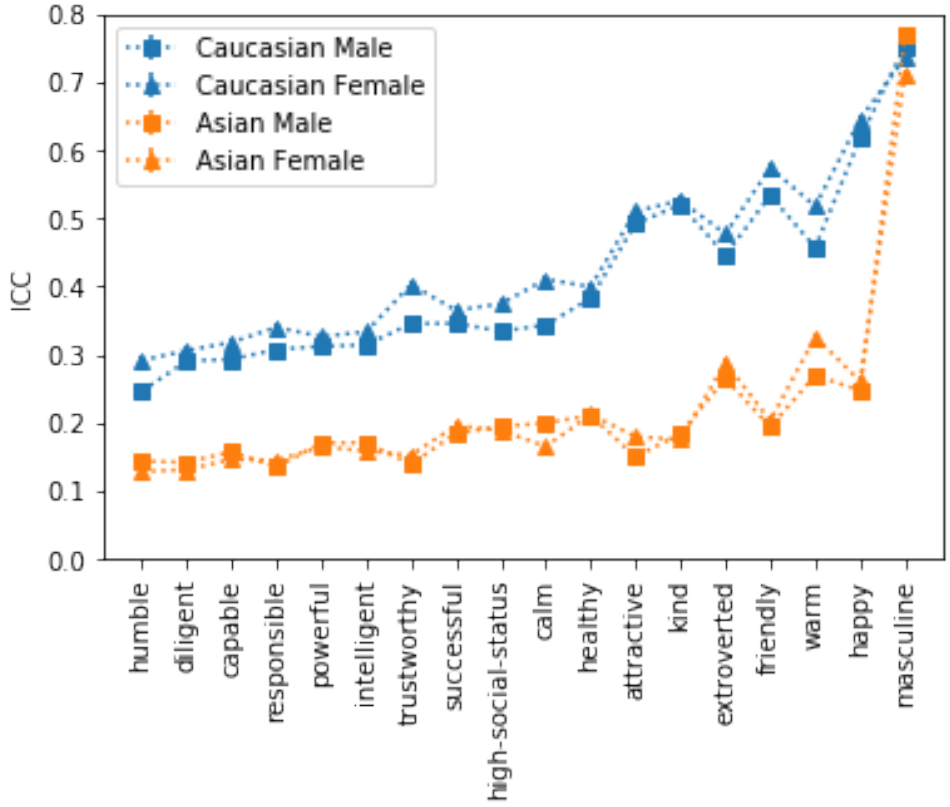


Figure 6.5. Raters are grouped along two dimensions: (1) gender (male/female), and (2) ethnicity (Asian/Caucasian). Traits are sorted from low to high based on average ICC. The warmth related traits also have high agreements, compared to competence related traits.

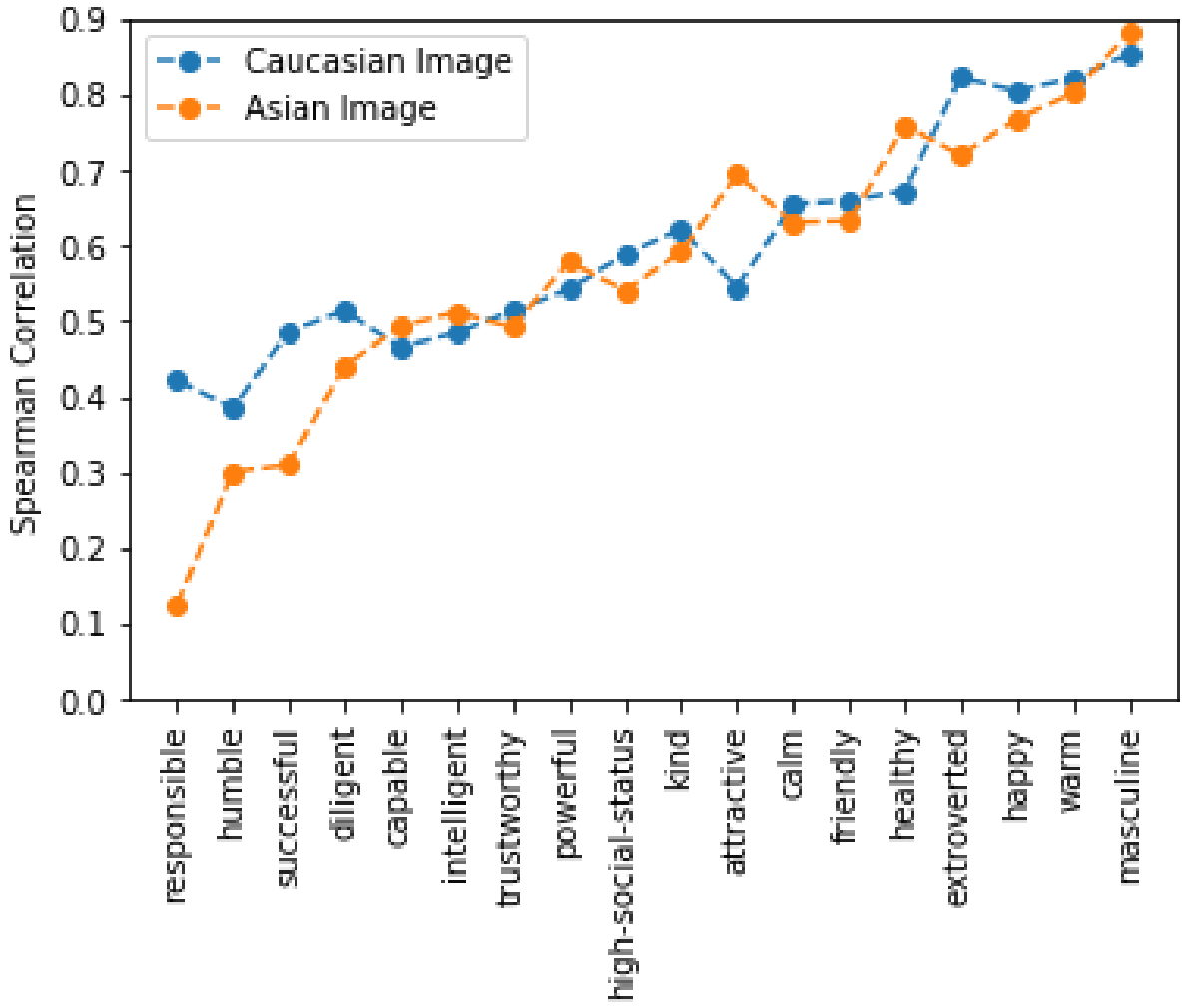


Figure 6.6. We split the images by the ethnicity, and plot the Spearman correlation to assess how Caucasian and Asian raters agree with each other on various traits, as well as the difference in agreement levels for Asian and Caucasian faces.

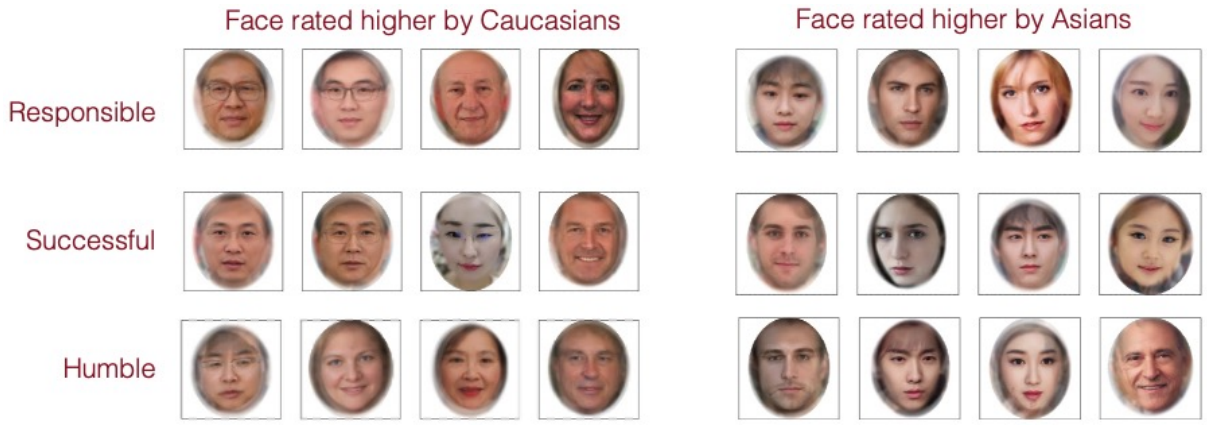


Figure 6.7. Morphed images that are rated most differently by Caucasians and Asians in responsible, successful and humble traits. Images on the left side are rated higher by Caucasians; images on the right side are rated higher by Asians.

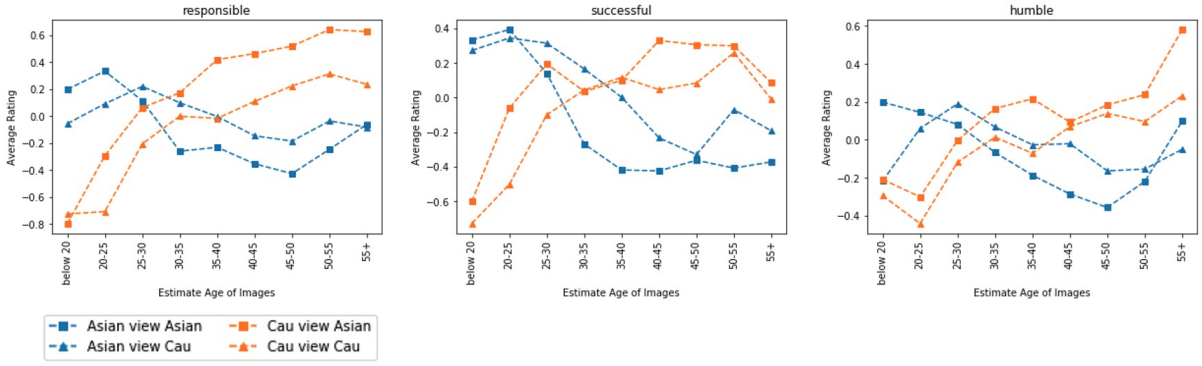


Figure 6.8. Faces are tagged into different age ranges, in five-year intervals. Blue lines are Asians' mean ratings and orange lines are Caucasians' mean ratings. Images are split by face ethnicity as well: square for Asian faces and triangle for Caucasian faces.

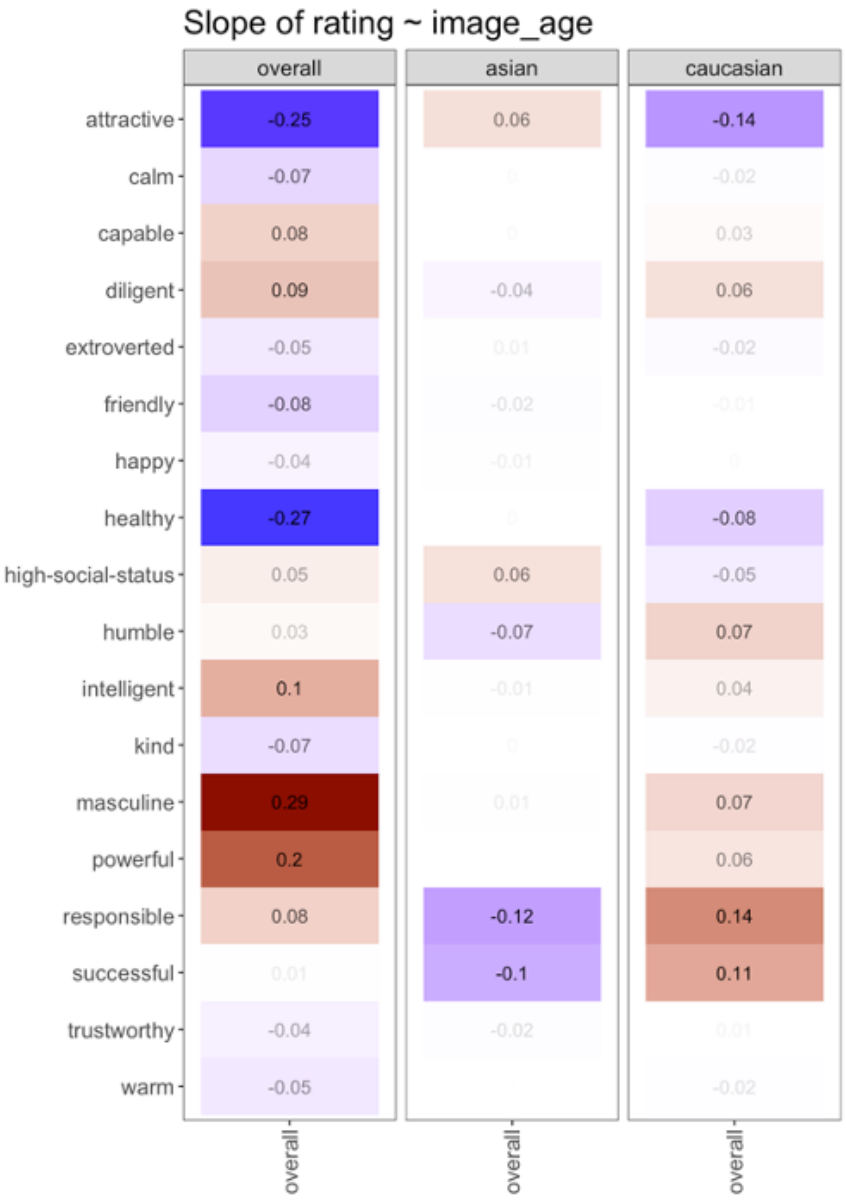


Figure 6.9. Slope of our linear regression model for each trait. The first column represents the overall trend with both Caucasian and Asian data. The values in the second and third columns are the magnitude of differences from the overall trend.

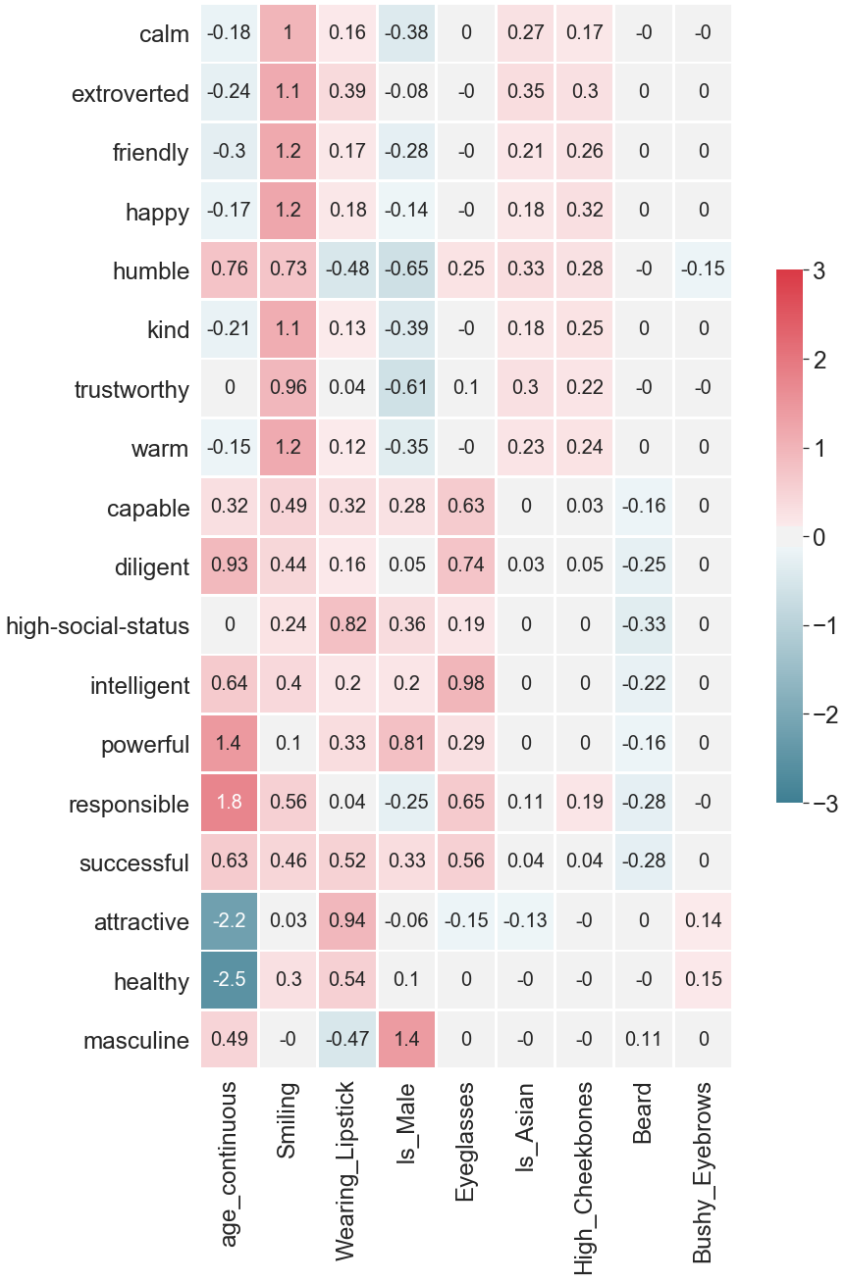


Figure 6.10. Heatmap for coefficients learned by a lasso regression model fit on Caucasian raters' data. The high level facial features (model inputs) and social impression traits (model outputs) are shown along the columns and rows respectively.

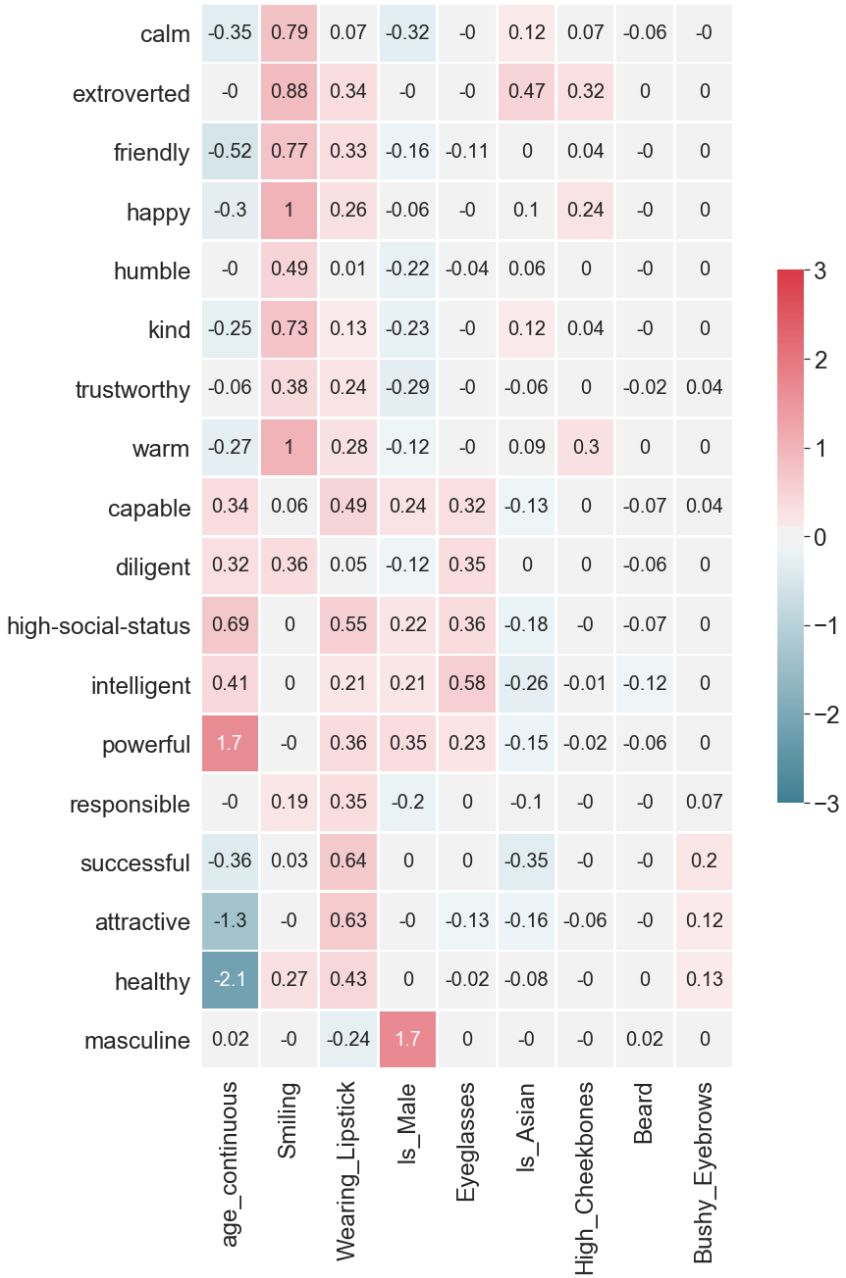


Figure 6.11. Heatmap for coefficients learned by a lasso regression model fit on Asian raters' data. The high level facial features (model inputs) and social impression traits (model outputs) are shown along the columns and rows respectively.

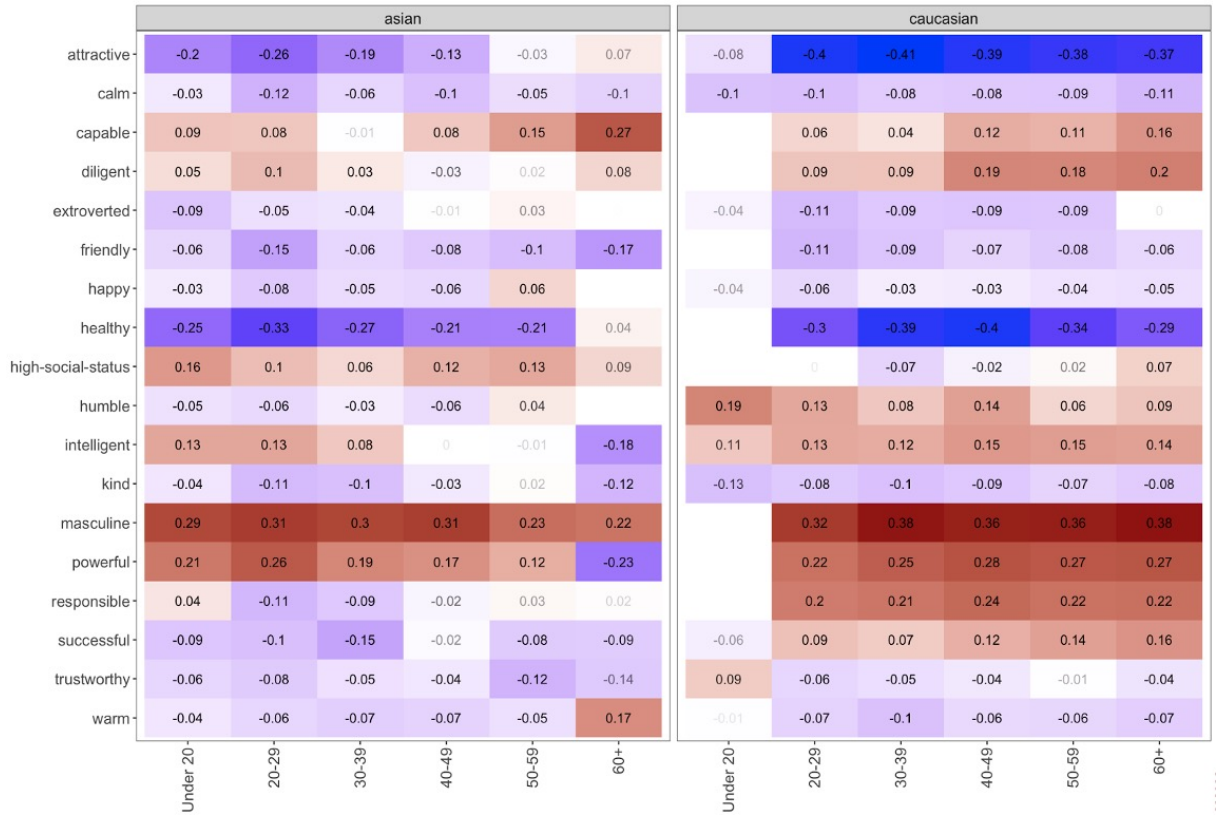


Figure 6.12. Rater are split into different age ranges, in ten-year intervals. We fit the slope of image age as a function of average ratings for each rater age-band. The left panel are Asian raters' results and the right panel are Caucasian raters' results.

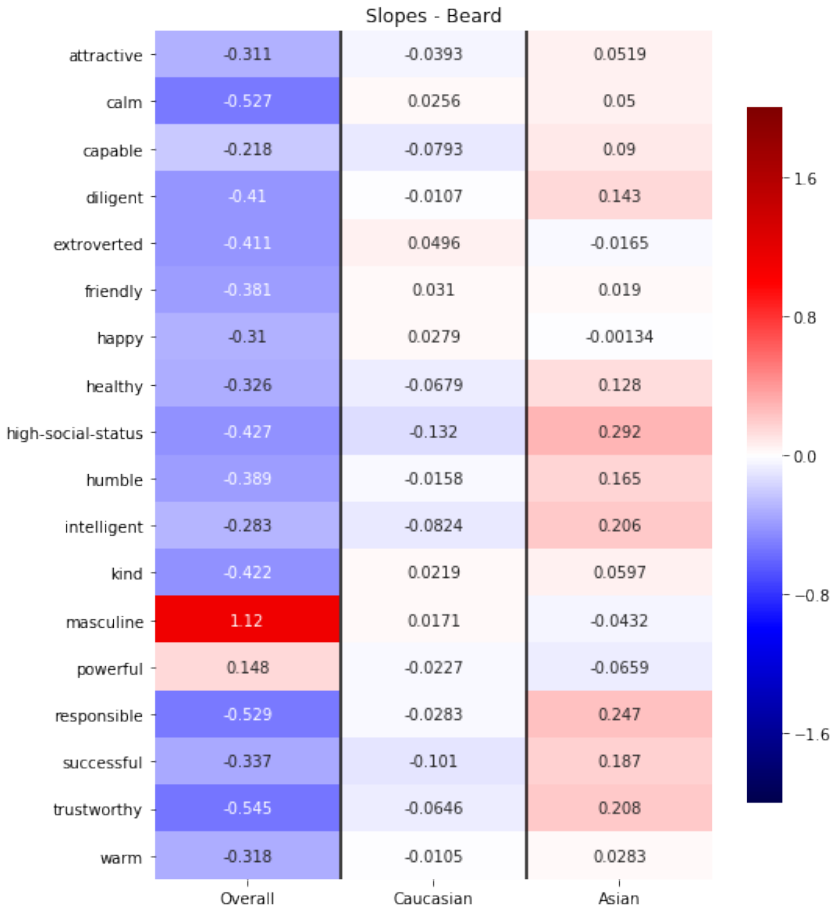


Figure 6.13. Slope of our univariate linear regression model for beard, the first column represents the overall trend with both Caucasian and Asian data. The values in the second and third columns are the magnitude of differences from the overall trend.

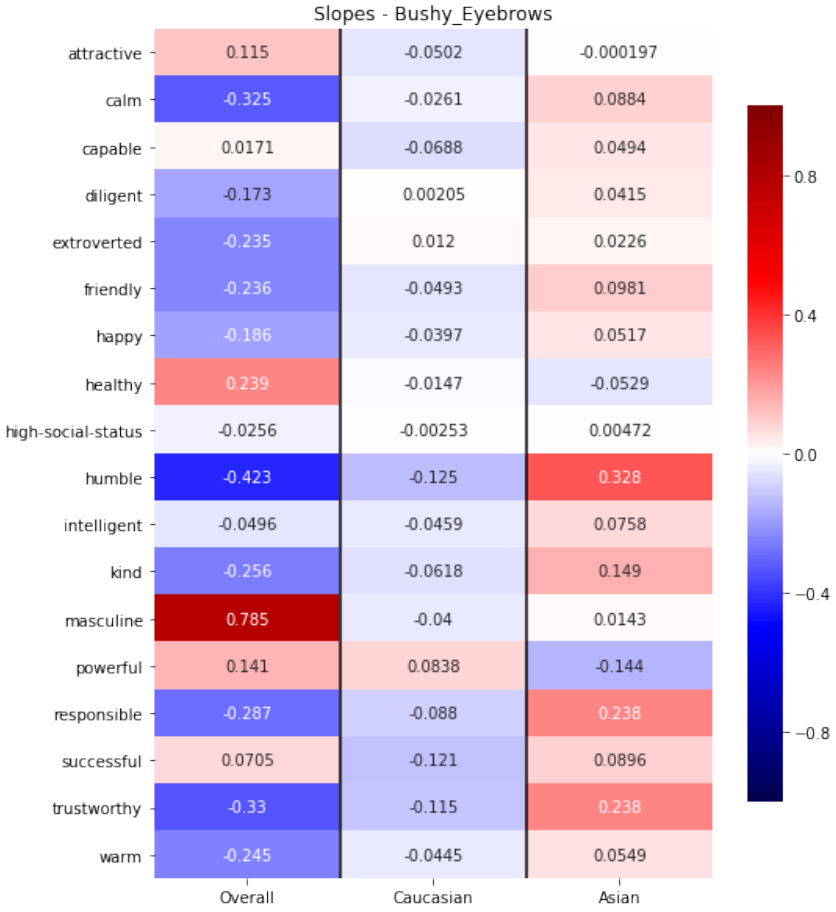


Figure 6.14. Slope of our univariate linear regression model for bushy eyebrow. The scaling color is encoded by magnitude of the coefficient (blue for negative values, red for positive values). The legend and color coding is the same from hereafter, except for the variable of interest.

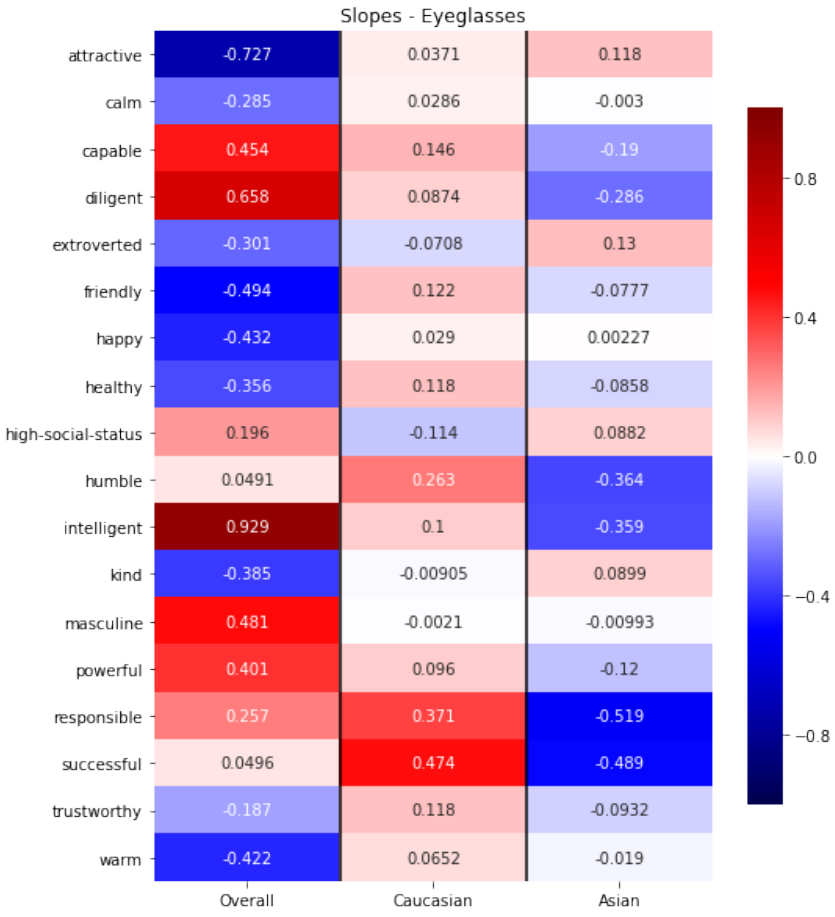


Figure 6.15. Slope of our univariate linear regression model for eyeglasses.

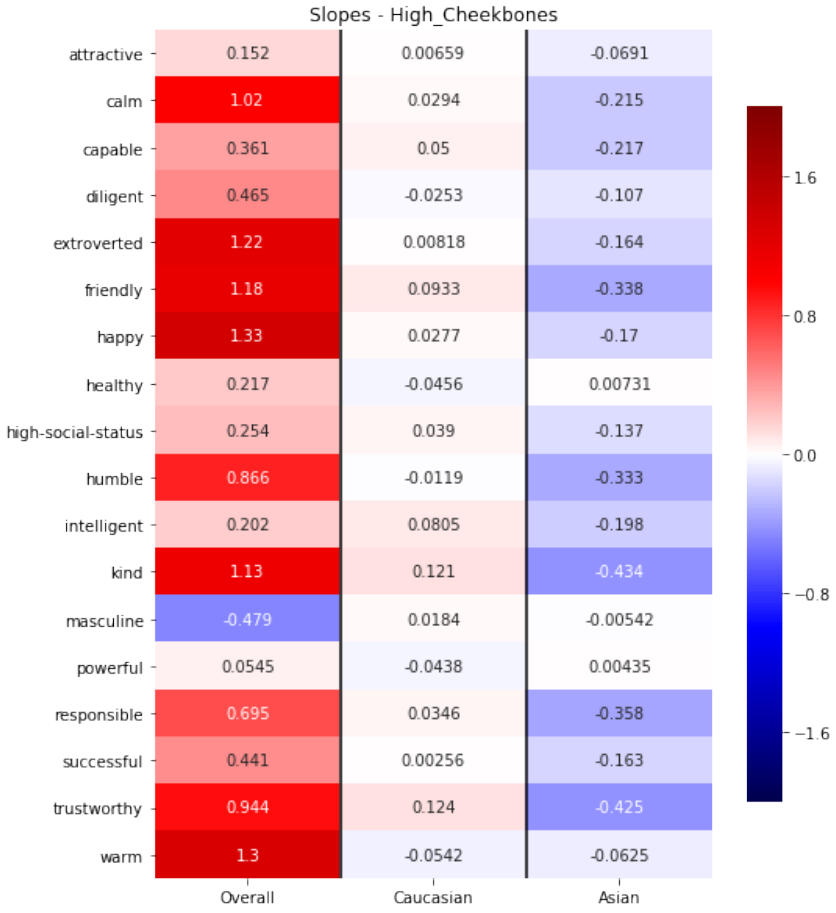


Figure 6.16. Slope of our univariate linear regression model for high cheekbones.

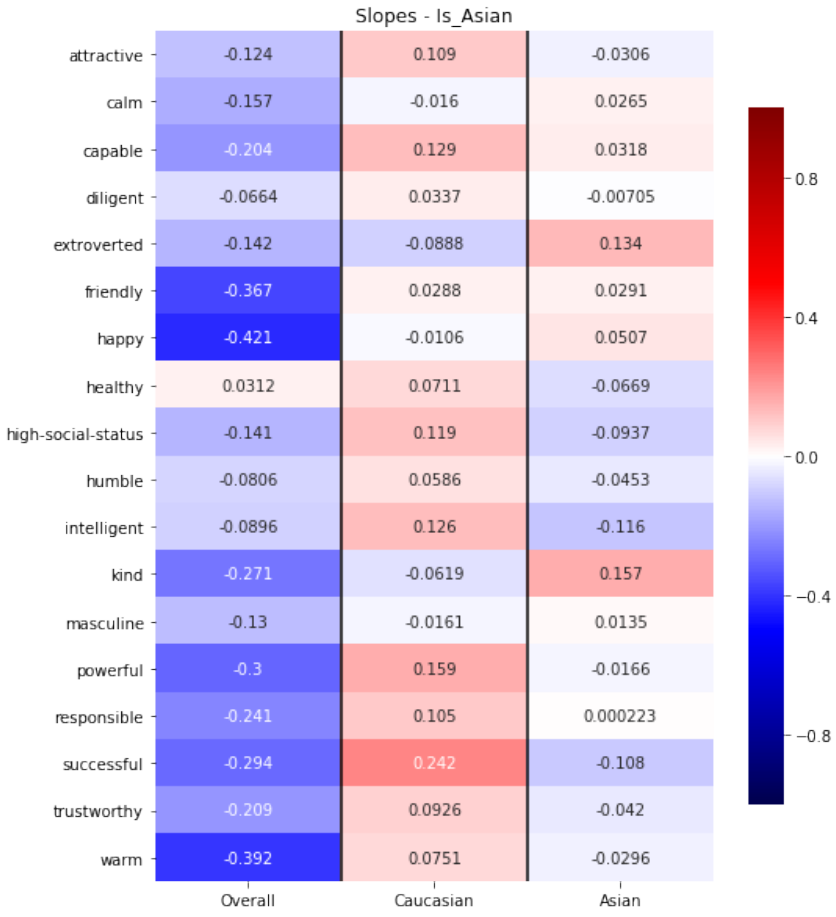


Figure 6.17. Slope of our univariate linear regression model for is-Asian.

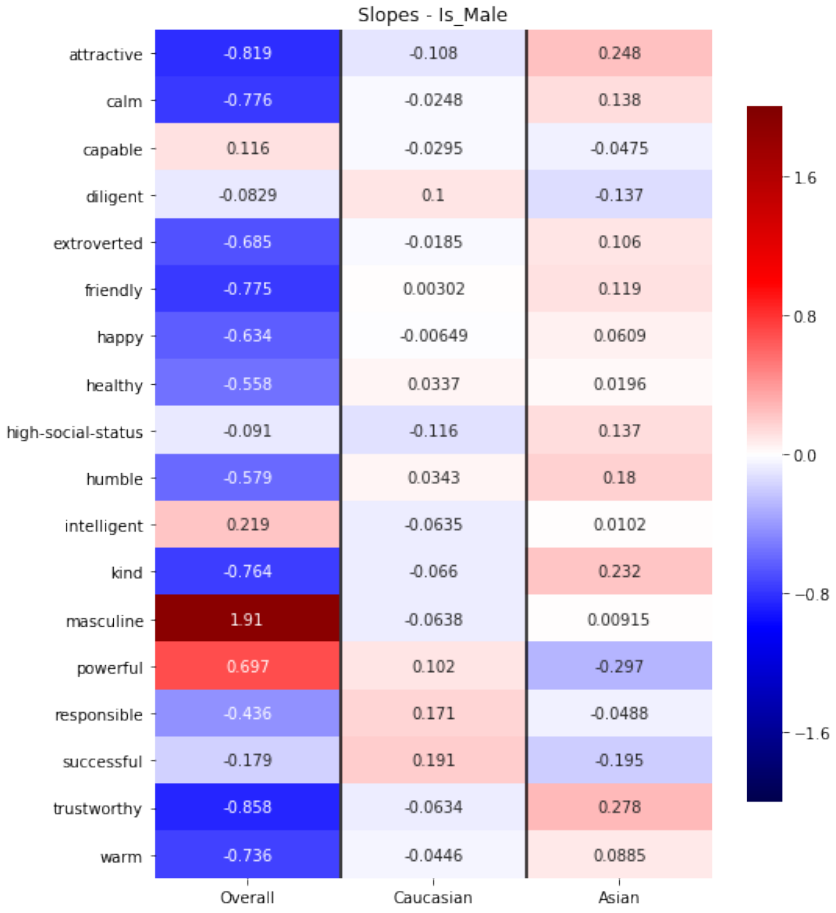


Figure 6.18. Slope of our univariate linear regression model for is-male.

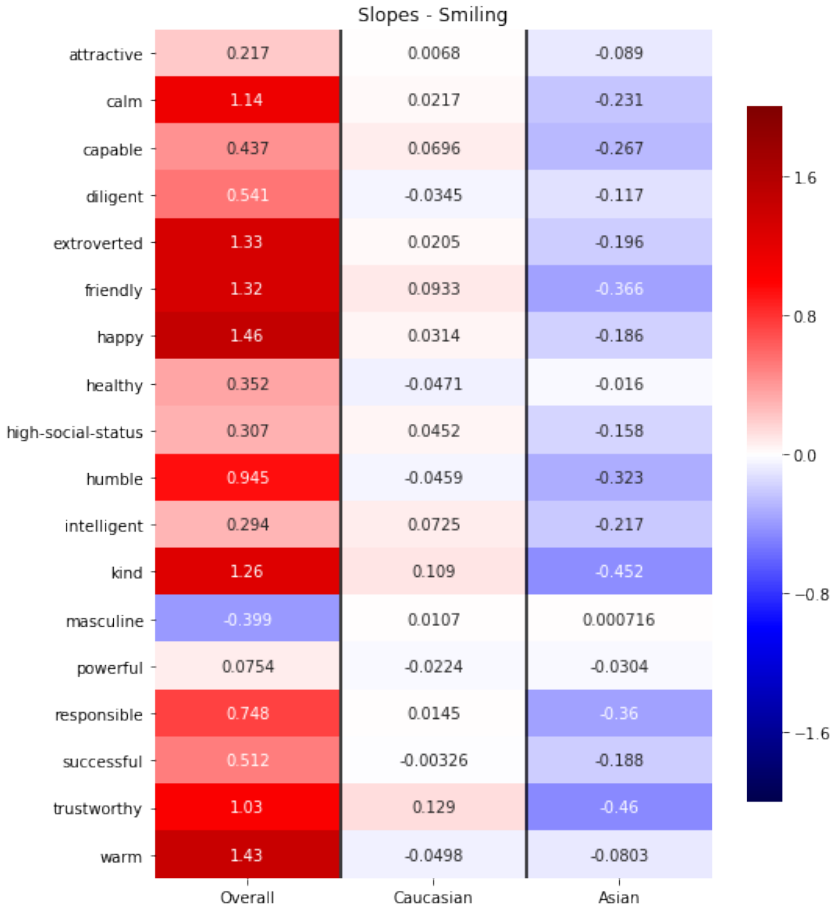


Figure 6.19. Slope of our univariate linear regression model for smiling.

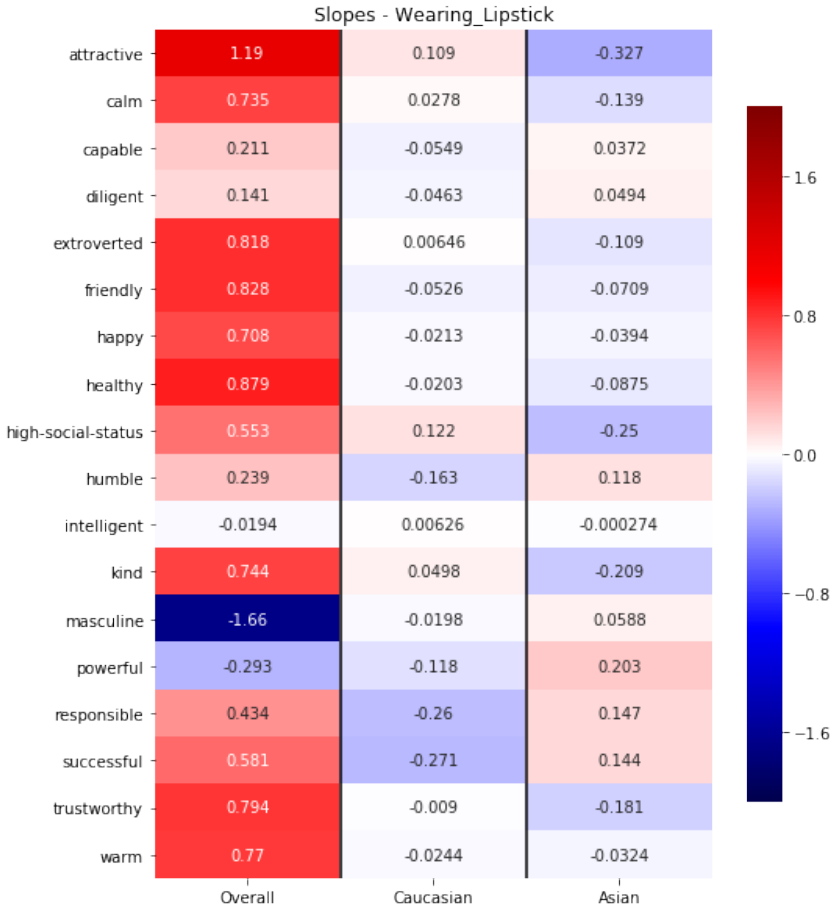


Figure 6.20. Slope of our univariate linear regression model for wearing lipstick.

Chapter 7

Conclusion

In this thesis, we examine social impressions of faces through the lens of computational modeling, psychological experiments, and cultural comparisons. First, we develop a model that automatically predicts human social impression judgments using neural networks. We learn that it's feasible to predict subjective impressions from objective images. And the more humans agree with each other, the better the model can learn to predict human responses.

Building on the predictive model, second, we build a generative model that can change faces holistically to augment or decrease specific characteristics, such as attractiveness and aggressiveness of a face. This modification model provides us a useful tool to traverse in impression space and go back to the image space to visualize the changes.

Third, we examine how specific physical attributes, such as hair color, affect impressions of faces, using a GAN model, and psychological experiments. This paradigm can be generalized to probe the causal relationship between any other physical facial features and impression traits.

Finally, we conduct a large-scale cross-cultural study, using 18 traits related to approachability, youthful-attractiveness, and competence evaluation, with Caucasian and Asian participants rating Caucasian and Asian faces. We investigate the mediating factors behind impression formation and estimate how high-level facial features such as age, eyeglasses, and smiles are judged similarly and differently by people from two cultures. We find that age is a cue that is used most differently across cultures. Senior people look more responsible, successful, diligent and humbler

in Caucasian people's eye. On the other hand, Asian observers rate as of higher social status, and yet less successful. Interestingly, we also find that Asian faces have a counter-productive effect on capability-related impressions, but only for Asian observers.

Overall, our work provides a computational framework to predict and modify faces, which is practically useful in real-life scenarios regarding optimal self-image presentation. Our model lays the foundation for the quantitative study of first impressions. Our psychological and cross-cultural studies reveal the universality and culture-specific judgments mediated by high-level features that affect first impressions. These findings motivate future research in social psychology to understand the deeper cultural roots behind the differences in first impression formation.

Our studies establish the link from face images to high-level facial features to social impression space and identify the unique role of observers' culture on impression formation. Future investigation on the deeper cultural root behind impression formation patterns would be fascinating.

Lastly, although it's human nature that we all want to make a good impression on others, it is crucial to be aware of the bias embedded behind our first impression formation and the dubious accuracy of these subjective impressions. Faces provide a shortcut to a lot of useful social information, but they don't offer a map to a person's true nature. Instead, the social impressions we form of others reveal our own personal history, cultural background, and biases. Our brains are prone to form impressions spontaneously by reading the face, but it requires us to use our rationality to cut the bias loop and make decisions wisely, not just at face value.

Bibliography

[1] Reginald B Adams Jr, Ursula Hess, and Robert E Kleck. The intersection of gender-related facial appearance and facial displays of emotion. *Emotion Review*, 7(1):5–13, 2015.

[2] Muhammad Aurangzeb Ahmad, Ankur Teredesai, and Carly Eckert. Fairness, accountability, transparency in ai at scale: lessons from national programs. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 690–690, 2020.

[3] Linda Albright, Thomas E Malloy, Qi Dong, David A Kenny, Xiaoyi Fang, Lynn Winquist, and Da Yu. Cross-cultural consensus in personality judgments. *Journal of personality and social psychology*, 72(3):558, 1997.

[4] Hani Altwaijry and Serge Belongie. Relative ranking of facial attractiveness. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 117–124. IEEE, 2013.

[5] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. *arXiv preprint arXiv:1702.01983*, 2017.

[6] John Antonakis and Olaf Dalgas. Predicting elections: Child's play! *Science*, 323(5918):1183–1183, 2009.

[7] Solomon E Asch. Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3):258, 1946.

[8] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4):1323, 2013.

[9] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4):1323, 2013.

[10] Moshe Bar, Maital Neta, and Heather Linz. Very first impressions. *Emotion*, 6(2):269, 2006.

[11] Susan A Basow. *Gender: Stereotypes and roles*. Thomson Brooks/Cole Publishing Co, 1992.

[12] Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.

[13] Diane S Berry and Leslie Z McArthur. Perceiving character in faces: the impact of age-related craniofacial changes on social perception. *Psychological bulletin*, 100(1):3, 1986.

[14] Diane S Berry and Leslie Zebrowitz-McArthur. What's in a face? facial maturity and the attribution of legal responsibility. *Personality and Social Psychology Bulletin*, 14(1):23–33, 1988.

[15] David Berthelot, Tom Schumm, and Luke Metz. BEGAN: boundary equilibrium generative adversarial networks. *CoRR*, abs/1703.10717, 2017.

[16] Caroline Blais, Rachael E Jack, Christoph Scheepers, Daniel Fiset, and Roberto Caldara. Culture shapes how we look at faces. *PloS one*, 3(8), 2008.

[17] Peter Borkenau and Anette Liebler. Trait inferences: Sources of validity at zero acquaintance. *Journal of personality and social psychology*, 62(4):645, 1992.

[18] Sheryl Brahnam. A computational model of the trait impressions of the face for agent perception and face synthesis. *AISB Journal*, 1(6):481–508, 2005.

[19] Vicki Bruce, A Mike Burton, Elias Hanna, Pat Healey, Oli Mason, Anne Coombes, Rick Fright, and Alf Linney. Sex discrimination: how do we tell the difference between male and female faces? *perception*, 22(2):131–152, 1993.

[20] A Mike Burton, Vicki Bruce, and Neal Dench. What's the difference between men and women? evidence from facial measurement. *Perception*, 22(2):153–176, 1993.

[21] A Mike Burton, Rob Jenkins, and Stefan R Schweinberger. Mental representations of familiar faces. *British Journal of Psychology*, 102(4):943–958, 2011.

[22] Justin M Carré, Mark D Morrissey, Catherine J Mondloch, and Cheryl M McCormick. Estimating aggression from emotionally neutral faces: Which facial cues are diagnostic? *Perception*, 39(3):356–377, 2010.

[23] Luigi Castelli, Luciana Carraro, Claudia Ghitti, and Massimiliano Pastore. The effects of perceived competence and sociability on electoral outcomes. *Journal of Experimental Social Psychology*, 45(5):1152–1155, 2009.

[24] Luke J Chang, Bradley B Doll, Mascha van't Wout, Michael J Frank, and Alan G Sanfey. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive psychology*, 61(2):87–105, 2010.

[25] Fang Fang Chen, Yiming Jing, and Jeong Min Lee. The looks of a leader: Competent and trustworthy, but not dominant. *Journal of Experimental Social Psychology*, 51:27–33, 2014.

[26] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR*, abs/1711.09020, 2017.

[27] Kaare Christensen, Mikael Thinggaard, Matt McGue, Helle Rexbye, Abraham Aviv, David Gunn, Frans van der Ouderaa, James W Vaupel, et al. Perceived age as clinically useful biomarker of ageing: cohort study. *Bmj*, 339:b5262, 2009.

[28] Glen U Cleeton and Frederick B Knight. Validity of character judgments based on external criteria. *Journal of Applied Psychology*, 8(2):215, 1924.

[29] Emily J Cogsdill, Alexander T Todorov, Elizabeth S Spelke, and Mahzarin R Banaji. Inferring character from faces: A developmental study. *Psychological science*, 25(5):1132–1139, 2014.

[30] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[31] G.W. Cottrell, P. Munro, and D. Zipser. Learning internal representations from gray-scale images: An example of extensional programming. In *Ninth Annual Conference of the Cognitive Science Society*, pages 462–473, Seattle 1987, 1987. Lawrence Erlbaum, Hillsdale.

[32] Antonia Creswell, Anil A. Bharath, and Biswa Sengupta. Conditional autoencoders with adversarial information factorization. *CoRR*, abs/1711.05175, 2017.

[33] Amy JC Cuddy, Susan T Fiske, Virginia SY Kwan, Peter Glick, Stephanie Demoulin, Jacques-Philippe Leyens, Michael Harris Bond, Jean-Claude Croizet, Naomi Ellemers, Ed Sleebos, et al. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, 48(1):1–33, 2009.

[34] Michael R Cunningham, Alan R Roberts, Anita P Barbee, Perri B Druen, and Cheng-Huan Wu. "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):261, 1995.

[35] Matthew N Dailey, Carrie Joyce, Michael J Lyons, Miyuki Kamachi, Hanae Ishi, Jiro Gyoba, and Garrison W Cottrell. Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion*, 10(6):874, 2010.

[36] Lisa M DeBruine. Facial resemblance enhances trust. *Proceedings of the Royal Society of London B: Biological Sciences*, 269(1498):1307–1312, 2002.

[37] Lisa M DeBruine. Facial resemblance increases the attractiveness of same–sex faces more than other–sex faces. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(1552):2085–2090, 2004.

[38] Lisa M DeBruine. Trustworthy but not lust-worthy: Context-specific effects of facial resemblance. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1566):919–922, 2005.

[39] David E. DeMers and Garrison W. Cottrell. Nonlinear dimensionality reduction. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 580–587, San Mateo, 1993. Morgan Kaufmann.

[40] Karen Dion, Ellen Berscheid, and Elaine Walster. What is beautiful is good. *Journal of personality and social psychology*, 24(3):285, 1972.

[41] Sonia Doallo, Jane E Raymond, Kimron L Shapiro, Monika Kiss, Martin Eimer, and Anna C Nobre. Response inhibition results in the emotional devaluation of faces: neural correlates as revealed by fmri. *Social Cognitive and Affective Neuroscience*, 7(6):649–659, 2011.

[42] Ron Dotsch and Alexander Todorov. Reverse correlating social face perception. *Social Psychological and Personality Science*, 3(5):562–571, 2012.

[43] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. Longitudinal face modeling via temporal deep restricted boltzmann machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5772–5780. IEEE Computer Society, 2016.

[44] Alice H Eagly, Richard D Ashmore, Mona G Makhijani, and Laura C Longo. What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological bulletin*, 110(1):109, 1991.

[45] Natalie C Ebner. Age of face matters: Age-group differences in ratings of young and old faces. *Behavior research methods*, 40(1):130–136, 2008.

[46] Catherine C Eckel and Ragan Petrie. Face value. *American Economic Review*, 101(4):1497–1513, 2011.

[47] Charles Efferson and Sonja Vogt. Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports*, 3:1047, 2013.

[48] Yael Eisenthal, Gideon Dror, and Eytan Ruppin. Facial attractiveness: Beauty and the machine. *Neural Computation*, 18(1):119–142, 2006.

[49] Paul Ekman, E Richard Sorenson, and Wallace V Friesen. Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88, 1969.

[50] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

[51] Virginia Falvello, Michael Vinson, Chiara Ferrari, and Alexander Todorov. The robustness of learning about the trustworthiness of other people. *Social Cognition*, 33(5):368, 2015.

[52] Susan T Fiske, Amy JC Cuddy, and Peter Glick. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83, 2007.

[53] Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition (2002). In *Social Cognition*, pages 171–222. Routledge, 2018.

[54] Jonathan B Freeman and Nalini Ambady. A dynamic interactive theory of person construal. *Psychological review*, 118(2):247, 2011.

[55] Jonathan B Freeman, Yina Ma, Shihui Han, and Nalini Ambady. Influences of culture and visual context on real-time social categorization. *Journal of experimental social psychology*, 49(2):206–210, 2013.

[56] Jonathan B Freeman, Andrew M Penner, Aliya Saperstein, Matthias Scheutz, and Nalini Ambady. Looking the part: Social status cues shape race perception. *PloS one*, 6(9):e25107, 2011.

[57] Heidi Friedman and Leslie A Zebrowitz. The contribution of typical sex differences in facial maturity to sex role stereotypes. *Personality and Social Psychology Bulletin*, 18(4):430–438, 1992.

[58] Jacob R. Gardner, Matt J. Kusner, Yixuan Li, Paul Upchurch, Kilian Q. Weinberger, and John E. Hopcroft. Deep manifold traversal: Changing labels with convolutional features. *CoRR*, abs/1511.06421, 2015.

[59] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[60] John R Graham, Campbell R Harvey, and Manju Puri. A corporate beauty contest. *Management Science*, 63(9):3044–3056, 2016.

[61] Douglas Gray, Kai Yu, Wei Xu, and Yihong Gong. Predicting facial beauty without landmarks. In *Computer Vision–ECCV 2010*, pages 434–447. Springer, 2010.

[62] Anthony G Greenwald and Linda Hamilton Krieger. Implicit bias: Scientific foundations. *California Law Review*, 94(4):945–967, 2006.

[63] Jinyan Guan, Chaitanya Ryali, and J Yu Angela. Computational modeling of social face perception in humans: Leveraging the active appearance model. *bioRxiv*, page 360776, 2018.

[64] Nicolas Guéguen. Hair color and courtship: Blond women received more courtship solicitations and redhead men received more refusals. *Psychological Studies*, 57(4):369–375, 2012.

[65] Gül Günaydin, Vivian Zayas, Emre Selcuk, and Cindy Hazan. I like you but i don't know why: Objective facial resemblance to significant others influences snap judgments. *Journal of Experimental Social Psychology*, 48(1):350–353, 2012.

[66] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008.

[67] Ran Hassin and Yaacov Trope. Facing faces: studies on the cognitive aspects of physiognomy. *Journal of personality and social psychology*, 78(5):837, 2000.

[68] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[69] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

[70] Eric Hehman, Clare AM Sutherland, Jessica K Flake, and Michael L Slepian. The unique contributions of perceiver and target characteristics in person perception. *Journal of personality and social psychology*, 113(4):513, 2017.

[71] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001.

[72] Ursula Hess, Sylvie Blairy, and Robert E Kleck. The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal behavior*, 24(4):265–283, 2000.

[73] James L Hilton and William Von Hippel. Stereotypes. *Annual review of psychology*, 47(1):237–271, 1996.

[74] Geoff E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[75] Geert Hofstede. Culture and organizations. *International Studies of Management & Organization*, 10(4):15–41, 1980.

[76] Ying Hu, Connor J Parde, Matthew Q Hill, Naureen Mahmood, and Alice J O'Toole. First impressions of personality traits from body shapes. *Psychological science*, page 0956797618799300, 2018.

[77] Roland Imhoff, Jonas Woelki, Sebastian Hanke, and Ron Dotsch. Warmth and competence in your face! visual encoding of stereotype content. *Frontiers in psychology*, 4:386, 2013.

[78] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[79] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[80] Tiffany A Ito, Eve C Willadsen-Jensen, Jesse T Kaye, and Bernadette Park. Contextual variation in automatic evaluative bias to racially-ambiguous faces. *Journal of experimental social psychology*, 47(4):818–823, 2011.

[81] Rob Jenkins, David White, Xandra Van Montfort, and A Mike Burton. Variability in photos of the same face. *Cognition*, 121(3):313–323, 2011.

[82] Kerri L Johnson, Jonathan B Freeman, and Kristin Pauker. Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of personality and social psychology*, 102(1):116, 2012.

[83] Benedict C Jones, Lisa Marie DeBruine, Jessica Kay Flake, Balazs Aczel, Matúš Adamkovič, Ravin Alaei, Sinan Alper, Michael R Andreychik, Daniel Ansari, Jack Dennis Arnal, et al. Social perception of faces around the world: How well does the valence-dominance model generalize across world regions?(registered report stage 1). 2018.

[84] Benedict C Jones, Anthony C Little, Ian S Penton-Voak, Bernard P Tiddeman, D Michael Burt, and David I Perrett. Facial symmetry and judgements of apparent health: support for a "good genes" explanation of the attractiveness–symmetry relationship. *Evolution and human behavior*, 22(6):417–429, 2001.

[85] Amit Kagian, Gideon Dror, Tommer Leyvand, Isaac Meilijson, Daniel Cohen-Or, and Eytan Ruppin. A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. *Vision Research*, 48(2):235–243, 2008.

[86] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[87] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[88] Ira Kemelmacher-Shlizerman. Transfiguring portraits. *ACM Trans. Graph.*, 35(4):94:1–94:8, 2016.

[89] Aditya Khosla, Wilma A Bainbridge, Antonio Torralba, and Aude Oliva. Modifying the memorability of face photographs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3200–3207, 2013.

[90] Aditya Khosla, Wilma A. Bainbridge, Antonio Torralba, and Aude Oliva. Modifying the memorability of face photographs. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 3200–3207. IEEE Computer Society, 2013.

[91] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[92] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[93] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[94] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[95] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[96] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[97] Kuba Krys, Karolina Hansen, Cai Xing, Piotr Szarota, and Miao-miao Yang. Do only fools smile at strangers? Cultural differences in social perception of intelligence of smiling individuals. *Journal of Cross-Cultural Psychology*, 45(2):314–321, 2014.

[98] Diana J Kyle and Heike IM Mahler. The effects of hair color and cosmetic use on perceptions of a female's ability. *Psychology of Women Quarterly*, 20(3):447–455, 1996.

[99] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, 33(4), 2014.

[100] Judith H Langlois and Lori A Roggman. Attractive faces are only average. *Psychological science*, 1(2):115–121, 1990.

[101] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[102] Aldo Laurentini and Andrea Bottino. Computer analysis of face beauty: A survey. *Computer Vision and Image Understanding*, 125:184–199, 2014.

[103] Chappell Lawson, Gabriel S Lenz, Andy Baker, and Michael Myers. Looking like a winner: Candidate appearance and electoral success in new democracies. *World Politics*, 62(4):561–593, 2010.

[104] Minhyeok Lee and Junhee Seok. Controllable generative adversarial network. *CoRR*, abs/1708.00598, 2017.

[105] Gloria Leventhal and Ronald Krate. Physical attractiveness and severity of sentencing. *Psychological Reports*, 40(1):315–318, 1977.

[106] Bochao D Lin, Gonneke Willemsen, Abdel Abdellaoui, Meike Bartels, Erik A Ehli, Gareth E Davies, Dorret I Boomsma, and Jouke J Hottenga. The genetic overlap between hair and eye color. *Twin Research and Human Genetics*, 19(6):595–599, 2016.

[107] Chujun Lin, Ralph Adolphs, and R Michael Alvarez. Inferring whether officials are corruptible from looking at their faces. *Psychological science*, page 0956797618788882, 2018.

[108] Walter Lippmann. Stereotypes. *Public Opinion and the Press. Nueva York: Mcmillan Publishing Co*, 1922.

[109] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. Turkprime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods*, 49, 04 2016.

[110] Anthony C Little and David I Perrett. Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology*, 98(1):111–126, 2007.

[111] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *CoRR*, abs/1703.00848, 2017.

[112] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[113] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[114] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[115] Robert W Livingston and Nicholas A Pearce. The teddy-bear effect: Does having a baby face benefit black chief executive officers? *Psychological science*, 20(10):1229–1236, 2009.

[116] G William Lucker, William E Beane, and Robert L Helmreich. The strength of the halo effect in physical attractiveness research. *The Journal of Psychology*, 107(1):69–75, 1981.

[117] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4):1122–1135, 2015.

[118] Kyle Mattes and Caitlin Milazzo. Pretty faces, marginal races: Predicting election outcomes using trait assessments of british parliamentary candidates. *Electoral Studies*, 34:177–189, 2014.

[119] Allan Mazur, Julie Mazur, and Caroline Keating. Military rank attainment of a west point class: Effects of cadets' physical features. *American Journal of Sociology*, 90(1):125–150, 1984.

[120] Leslie Z McArthur and Reuben M Baron. Toward an ecological theory of social perception. *Psychological review*, 90(3):215, 1983.

[121] Leslie Zebrowitz McArthur and Diane S Berry. Cross-cultural agreement in perceptions of babyfaced adults. *Journal of cross-cultural psychology*, 18(2):165–192, 1987.

[122] Mel McCurrie, Fernando Beletti, Lucas Parzianello, Allen Westendorp, Samuel Anthony, and Walter J Scheirer. Predicting first impressions with deep learning. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 518–525. IEEE, 2017.

[123] Alain Mignault and Avi Chaudhuri. The many faces of a neutral face: Head tilt and perception of dominance and emotion. *Journal of Nonverbal Behavior*, 27(2):111–132, 2003.

[124] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[125] Joann M Montepare and Heidi Dobish. The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal behavior*, 27(4):237–254, 2003.

[126] Ulrich Mueller and Allan Mazur. Facial dominance of west point cadets as a predictor of later military rank. *Social forces*, 74(3):823–850, 1996.

[127] Richard E Nisbett and Yuri Miyamoto. The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences*, 9(10):467–473, 2005.

[128] Ryo Oda, Noriko Yamagata, Yuki Yabiku, and Akiko Matsumoto-Oda. Altruism can be assessed correctly based on impression. *Human Nature*, 20(3):331–341, 2009.

[129] Christopher Y Olivola, Abigail B Sussman, Konstantinos Tsetsos, Olivia E Kang, and Alexander Todorov. Republicans prefer republican-looking leaders: Political facial stereotypes predict candidate electoral success among right-leaning voters. *Social Psychological and Personality Science*, 3(5):605–613, 2012.

[130] Christopher Y Olivola and Alexander Todorov. Fooled by first impressions? reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, 46(2):315–324, 2010.

[131] CY Olivola, PW Eastwick, EJ Finkel, A Hortaçsu, D Ariely, and A Todorov. A picture is worth a thousand inferences: First impressions and mate selection in internet matchmaking and speed-dating. Technical report, Working paper, Tepper School of Business, Carnegie Mellon University.[CYO], 2016.

[132] Nikolaas N Oosterhof and Alexander Todorov. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105(32):11087–11092, 2008.

[133] Daphna Oyserman, Heather M Coon, and Markus Kemmelmeier. Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. *Psychological bulletin*, 128(1):3, 2002.

[134] K. Hornik P. Baldi. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1988.

[135] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[136] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[137] Ian S Penton-Voak, Nicholas Pound, Anthony C Little, and David I Perrett. Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social cognition*, 24(5):607–640, 2006.

[138] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. Invertible conditional gans for image editing. *CoRR*, abs/1611.06355, 2016.

[139] Stephen Porter, Leanne ten Brinke, and Chantal Gustaw. Dangerous decisions: The impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime & Law*, 16(6):477–491, 2010.

[140] Rolf Reber, Norbert Schwarz, and Piotr Winkielman. Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and social psychology review*, 8(4):364–382, 2004.

[141] Lawrence Ian Reed, Katharine N Zeglen, and Karen L Schmidt. Facial expressions as honest signals of cooperative intent in a one-shot anonymous prisoner's dilemma game. *Evolution and Human Behavior*, 33(3):200–209, 2012.

[142] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic back-propagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[143] Constantin Rezlescu, Brad Duchaine, Christopher Y Olivola, and Nick Chater. Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PloS one*, 7(3):e34293, 2012.

[144] Gillian Rhodes, Linda Jeffery, Tamara L Watson, Colin WG Clifford, and Ken Nakayama. Fitting the mind to the world: Face adaptation and attractiveness aftereffects. *Psychological science*, 14(6):558–566, 2003.

[145] Gillian Rhodes, Fiona Proffitt, Jonathon M Grady, and Alex Sumich. Facial symmetry and the perception of beauty. *Psychonomic Bulletin & Review*, 5(4):659–669, 1998.

[146] Gillian Rhodes and Tanya Tremewan. Averageness, exaggeration, and facial attractiveness. *Psychological science*, 7(2):105–110, 1996.

[147] Karolann Robinson, Caroline Blais, Justin Duncan, Hélène Forget, and Daniel Fiset. The dual nature of the human face: there is a little jekyll and a little hyde in all of us. *Frontiers in psychology*, 5:139, 2014.

[148] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Some like it hot-visual guidance for preference prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5553–5561, 2016.

[149] Nicholas O Rule and Nalini Ambady. She's got the look: Inferences from female chief executive officers' faces predict their success. *Sex Roles*, 61(9-10):644–652, 2009.

[150] Nicholas O Rule and Nalini Ambady. Democrats and republicans can be differentiated from their faces. *PloS one*, 5(1):e8733, 2010.

[151] Nicholas O Rule, Nalini Ambady, and Katherine C Hallett. Female sexual orientation is perceived accurately, rapidly, and automatically from the face and its features. *Journal of Experimental Social Psychology*, 45(6):1245–1251, 2009.

[152] Nicholas O Rule, James V Garrett, and Nalini Ambady. On the perception of religious group membership from faces. *PloS one*, 5(12):e14241, 2010.

[153] Nicholas O Rule, Keiko Ishii, and Nalini Ambady. Cross-cultural impressions of leaders' faces: Consensus and predictive validity. *International Journal of Intercultural Relations*, 35(6):833–841, 2011.

[154] Nicholas O Rule, Keiko Ishii, Nalini Ambady, Katherine S Rosen, and Katherine C Hallett. Found in translation: Cross-cultural consensus in the accurate categorization of male sexual orientation. *Personality and Social Psychology Bulletin*, 37(11):1499–1507, 2011.

[155] Nicholas O Rule, Anne C Krendl, Zorana Ivcevic, and Nalini Ambady. Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology*, 104(3):409, 2013.

[156] James A Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin*, 115(1):102, 1994.

[157] Christopher P Said, Nicu Sebe, and Alexander Todorov. Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, 9(2):260, 2009.

[158] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[159] Joanna E Scheib, Steven W Gangestad, and Randy Thornhill. Facial attractiveness, symmetry and cues of good genes. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1431):1913–1917, 1999.

[160] Erik J Schlicht, Shinsuke Shimojo, Colin F Camerer, Peter Battaglia, and Ken Nakayama. Human wagering behavior depends on opponents' faces. *PloS one*, 5(7):e11663, 2010.

[161] Paul F Secord, William F Dukes, and William Bevan. Personalities in faces: I. an experiment in social perceiving. *Genetic psychology monographs*, 1954.

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

[163] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[164] Tamara Sims, Jeanne L Tsai, Da Jiang, Yaheng Wang, Helene H Fung, and Xiulan Zhang. Wanting to maximize the positive and minimize the negative: Implications for mixed affective experience in american and chinese contexts. *Journal of Personality and Social Psychology*, 109(2):292, 2015.

[165] Alan Slater, Gavin Bremner, Scott P Johnson, Penny Sherwood, Rachel Hayes, and Elizabeth Brown. Newborn infants' preference for attractive faces: The role of internal and external facial features. *Infancy*, 1(2):265–274, 2000.

[166] Carmel Sofer, Ron Dotsch, Masanori Oikawa, Haruka Oikawa, Daniel HJ Wigboldus, and Alexander Todorov. For your local eyes only: culture-specific face typicality influences perceptions of trustworthiness. *Perception*, 46(8):914–928, 2017.

[167] Carmel Sofer, Ron Dotsch, Daniel HJ Wigboldus, and Alexander Todorov. What is typical is good: The influence of face typicality on perceived trustworthiness. *Psychological Science*, 26(1):39–47, 2015.

[168] Amanda Song, Linjie Li, Chad Atalla, and Garrison Cottrell. Learning to see people like people. *arXiv preprint arXiv:1705.04282*, 2017.

[169] Adam Sparks, Tyler Burleigh, and Pat Barclay. We can see inside: Accurate prediction of prisoner's dilemma decisions in announced games following a face-to-face interaction. *Evolution and Human Behavior*, 37(3):210–216, 2016.

[170] Damian A Stanley, Peter Sokol-Hessner, Mahzarin R Banaji, and Elizabeth A Phelps. Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, 108(19):7710–7715, 2011.

[171] Russell Stewart, Mykhaylo Andriluka, and Andrew Y. Ng. End-to-end people detection in crowded scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2016)*, June 2016.

[172] Michael Stirrat and David I Perrett. Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological science*, 21(3):349–354, 2010.

[173] Clare Sutherland, Xixi Liu, Ying Chu, Lingshan Zhang, Julian Oldmeadow, and Andrew Young. Chinese perceivers' facial first impressions. *Journal of vision*, 15(12):1218–1218, 2015.

[174] Clare AM Sutherland, Xizi Liu, Lingshan Zhang, Yingtung Chu, Julian A Oldmeadow, and Andrew W Young. Facial first impressions across culture: Data-driven modeling of chinese and british perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin*, 44(4):521–537, 2018.

[175] Clare AM Sutherland, Julian A Oldmeadow, Isabel M Santos, John Towler, D Michael Burt, and Andrew W Young. Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, 127(1):105–118, 2013.

[176] Viren Swami, Adrian Furnham, and Kiran Joshi. The influence of skin tone, hair length, and hair colour on ratings of women's physical attractiveness, health and fertility. *Scandinavian Journal of Psychology*, 49(5):429–437, 2008.

[177] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2015)*, pages 1–9, 2015.

[178] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2014)*, pages 1701–1708, 2014.

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

[180] Dustin Tingley. Face-off: Facial features and strategic choice. *Political Psychology*, 35(1):35–55, 2014.

[181] Alexander Todorov, Sean G Baron, and Nikolaas N Oosterhof. Evaluating face trustworthiness: a model based approach. *Social Cognitive and Affective Neuroscience*, 3(2):119–127, 2008.

[182] Alexander Todorov, Ron Dotsch, Jenny M Porter, Nikolaas N Oosterhof, and Virginia B Falvello. Validation of data-driven computational models of social perception of faces. *Emotion*, 13(4):724, 2013.

[183] Alexander Todorov, Anesu N Mandisodza, Amir Goren, and Crystal C Hall. Inferences of competence from faces predict election outcomes. *Science*, 308(5728):1623–1626, 2005.

[184] Alexander Todorov, Christopher Y Olivola, Ron Dotsch, and Peter Mende-Siedlecki. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual review of psychology*, 66:519–545, 2015.

[185] Alexander Todorov, Manish Pakrashi, and Nikolaas N Oosterhof. Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, 27(6):813–833, 2009.

[186] Alexander Todorov and Jenny M Porter. Misleading first impressions: Different for different facial images of the same person. *Psychological science*, 25(7):1404–1417, 2014.

[187] Alexander Todorov, Chris P. Said, Andrew D. Engell, and Nikolaas N. Oosterhof. Understanding evaluation of faces on social dimensions. *Trends in cognitive sciences*, 12 12:455–60, 2008.

[188] Alexander T Todorov, Christopher C Said, and Sara C Verosky. Personality impressions from facial appearance. In *Oxford Handbook of Face Perception*. Oxford University Press, 2012.

[189] Zhuowen Tu. Learning generative models via discriminative approaches. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[190] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.

[191] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7064–7073, 2017.

[192] Jeffrey M Valla, Stephen J Ceci, and Wendy M Williams. The accuracy of inferences about criminality based on facial appearance. *Journal of Social, Evolutionary, and Cultural Psychology*, 5(1):66, 2011.

[193] Mascha Van't Wout and Alan G Sanfey. Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, 108(3):796–803, 2008.

[194] Richard JW Vernon, Clare AM Sutherland, Andrew W Young, and Tom Hartley. Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, 111(32):E3353–E3361, 2014.

[195] Jan Verplaetse, Sven Vanneste, and Johan Braeckman. You can judge a book by its cover: the sequel.: A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, 28(4):260–271, 2007.

[196] Sonja Vogt, Charles Efferson, and Ernst Fehr. Can we see inside? predicting strategic behavior given limited information. *Evolution and Human Behavior*, 34(4):258–264, 2013.

[197] Mirella Walker and Thomas Vetter. Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, 9(11):12–12, 2009.

[198] Michael A Webster and Otto H Maclin. Figural aftereffects in the perception of faces. *Psychonomic bulletin & review*, 6(4):647–653, 1999.

[199] Susan Weir and Margret Fine-Davis. 'dumb blonde'and 'temperamental redhead': The effect of hair colour on some attributed personality characteristics of women. *The Irish Journal of Psychology*, 10(1):11–19, 1989.

[200] Janine Willis and Alexander Todorov. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592–598, 2006.

[201] John Paul Wilson and Nicholas O Rule. Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological science*, 26(8):1325–1331, 2015.

[202] Piotr Winkielman and John T Cacioppo. Mind at ease puts a smile on the face: psychophysiological evidence that processing facilitation elicits positive affect. *Journal of personality and social psychology*, 81(6):989, 2001.

[203] Piotr Winkielman, Jamin Halberstadt, Tedra Fazendeiro, and Steve Catty. Prototypes are attractive because they are easy on the mind. *Psychological science*, 17(9):799–806, 2006.

[204] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[205] Weidong Yin, Yanwei Fu, Leonid Sigal, and Xiangyang Xue. Semi-latent GAN: learning to generate and modify facial images from attributes. *CoRR*, abs/1704.02166, 2017.

[206] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[207] Alison I Young, Kyle G Ratner, and Russell H Fazio. Political attitudes bias the mental representation of a presidential candidate's face. *Psychological Science*, 25(2):503–510, 2014.

[208] Leslie A Zebrowitz, P Matthew Bronstad, and Hoon Koo Lee. The contribution of face familiarity to ingroup favoritism and stereotyping. *Social Cognition*, 25(2):306–338, 2007.

[209] Leslie A Zebrowitz, Judith A Hall, Nora A Murphy, and Gillian Rhodes. Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin*, 28(2):238–249, 2002.

[210] Leslie A Zebrowitz, Masako Kikuchi, and Jean-Marc Fellous. Facial resemblance to emotions: group differences, impression effects, and race stereotypes. *Journal of personality and social psychology*, 98(2):175, 2010.

[211] Leslie A Zebrowitz and Joann M Montepare. Impressions of babyfaced individuals across the life span. *Developmental psychology*, 28(6):1143, 1992.

[212] Leslie A Zebrowitz and Gillian Rhodes. Sensitivity to "bad genes" and the anomalous face overgeneralization effect: Cue validity, cue utilization, and accuracy in judging intelligence and health. *Journal of nonverbal behavior*, 28(3):167–185, 2004.

[213] Leslie A Zebrowitz, Ruoxue Wang, P Matthew Bronstad, Dan Eisenberg, Eduardo Undurraga, Victoria Reyes-García, and Ricardo Godoy. First impressions from faces among us and culturally isolated tsimane'people in the bolivian rainforest. *Journal of Cross-Cultural Psychology*, 43(1):119–134, 2012.

[214] Leslie A Zebrowitz, Benjamin White, and Kristin Wieneke. Mere exposure and racial prejudice: Exposure to other-race faces increases liking for strangers of that race. *Social cognition*, 26(3):259–275, 2008.

[215] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV-2014)*, pages 818–833. Springer, 2014.

[216] Yang Zhong, Josephine Sullivan, and Haibo Li. Face attribute prediction using off-the-shelf cnn features. In *2016 International Conference on Biometrics (ICB)*, pages 1–7. IEEE, 2016.

[217] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[218] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[219] Daniel Zwillinger and Stephen Kokoska. *Standard Probability and Statistics Tables and Formulae*. Chapman & Hall/CRC, 2000.