INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Trustworthy Facial Analysis with Scarce and Biased Data

by

Shruti Nagpal

Under the supervision of

Dr. Mayank Vatsa

Dr. Richa Singh

Indraprastha Institute of Information Technology Delhi

September, 2022

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY DELHI

Trustworthy Facial Analysis with Scarce and Biased Data

by

Shruti Nagpal

Submitted
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

to the
Indraprastha Institute of Information Technology Delhi
September, 2022

Certificate

This is to certify that the thesis titled "**Trustworthy Facial Analysis with Scarce and Biased Data**" being submitted by **Shruti Nagpal** (PhD15002) to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

September, 2022
Dr. Mayank Vatsa
Professor

September, 2022
Dr. Richa Singh
Professor

Indraprastha Institute of Information Technology Delhi
New Delhi

Acknowledgment

My PhD journey has been more than just about research, innovation and pushing boundaries. It has been a journey of discovery, perseverance and breaking glass ceilings, and has been nothing short of a rollercoaster ride. It would not have been the same without the mentorship, guidance, love, support, and kindness I received from some of the phenomenal people during this very crucial chapter of my life.

First and foremost, I would like to express my heartfelt gratitude to my advisors and mentors Prof. Mayank Vatsa and Prof. Richa Singh for believing in my strengths. They taught me to stand for myself and believe in myself despite circumstances and situations. Thank you for the ghar wali chai when I was feeling lost, for letting us use the white board in your offices to explore new research avenues and navigate through bottlenecks while listening patiently. I would also like to thank Prof. Afzel Noore for hosting me during my internship at West Virginia University. His ability to listen patiently and think out of the box inspired me to be a better researcher. I am also grateful to Dr. Nalini Ratha and Dr. Angshul Majumdar for their technical expertise, feedback, and prompt responses on our collaboration projects. I also appreciate the yearly feedback I received from Dr. Saket Anand and Dr. Ganesh Bagler through the years. I am sincerely grateful to Dr. Pankaj Jalote for enabling me to roll over from the B.Tech program to the PhD program at IIIT Delhi and giving the opportunity and research freedom at par with some of the leading universities of the world. A big thank you to Priti ma'am, Sheetu ma'am, and Ashutosh sir for making everything on the administrative front so effortless and smooth. Thank you for putting up with last minute requests, changes, a zillion questions posed, and providing the best possible solutions. I would also like to thank IIIT-Delhi for being my second home for the past several years. It has given me some of the best memories, friendships, collaborations and so much more. It has shaped me into who I am today. The support from administration, faculty, the facilities and work environment made this journey so much smoother. I would also like to thank TCS and Facebook for their financial support for funding my research. I would also like to express my gratitude to each of my collaborators: Maneet, Naman, Daksha, Akshay, Soumyadeep, Isha, Arushi, Sanchit Gupta, Sanchit Sinha, Nikita, and Mohit Agarwal for all the discussions, and brainstorming sessions which made work more interesting and taught me something new. A special mention to friends away from

home: Dr. Naman and Dr. Daksha, who taught me the way around in a new country, pampered me during our time in the US, and were always up to try a new restaurant.

I would not be where I am without the unconditional support of my family and friends, who have been my pillars of strength and loudest cheerleaders since forever. My parents: Neeta and Naresh who inspired me to aim for the sky. They taught me that nothing is impossible to achieve and failures should be merely treated as stepping stones in the journey. More importantly, they taught me the importance of taking a break and enjoying the little things during the hustle. Thank you for all the times you drove halfway across the city to ensure I took a break and literally brought home to me, be it after a hectic day at work for a casual dinner or on a Sunday afternoon with home cooked food which was enjoyed in the campus lawns filled with warmth and laughter. My little brother, Uday, who is not so little anymore, for always having my back and checking in on me subtly with random song suggestions, unending discussions on food and healthy eating, and reminding me to "chill!". Maneet Singh for being my research partner, pseudo room-mate, support system, best friend, family away from home. She inspires me to be better everyday. This roller coaster ride would not have been possible without her support, and our unsaid rituals of GK market visits for momos and cold coffee to find motivation on a slow day, victory dances and Social visits for big and small wins alike, Sushi order-ins and staircase rants after rejections. My dearest friends: Ankita, Jayasi, Priya, Geet and Hada who always had my back in their own way, be it to listen to me rant, feeding me with unending amounts of cold coffee, or dropping by on campus for a walk and biryani or game of Basketball. Lastly, I would also like to thank Dr. Mahalaxmi for teaching me to be kind to myself.

My heartfelt gratitude to each and every person mentioned above. Each person has influenced me and helped shape this dissertation in their own unique way. I hope to create a positive impact with my work in the years to come.

Shruti Nagpal

Trustworthy Facial Analysis with Scarce and Biased Data

by

Shruti Nagpal

Abstract

Face is one of the least invasive biometric modalities, and has been used as a physical signature to perform person recognition. Face recognition has widespread applicability in various domains such as surveillance, access control, and social media tagging. While face recognition has achieved very high performance in some settings, newer challenges such as developing trustworthy AI systems have emerged. Trustworthy facial analysis relies on three components: data, algorithm, and deployment. This dissertation focuses on the data centric challenges, specifically developing facial analysis models with scarce and biased data. Limited attention has been given to applications with the availability of scarce data, particularly heterogeneous data, i.e., data belonging to different domains, such as sketch to digital face image matching. Such applications have societal impact, but are often challenging in nature. With the rapid increase in number of automated systems, recently, the biased nature of facial analysis systems has also been highlighted, demonstrating the need for automated systems to be fair. There is growing literature of research being done to understand this challenge along with efforts to ensure that these systems are unbiased and work equally well for all sub-groups of our society, irrespective of gender, ethnicity, age, or demographics. To this effect, this dissertation focuses on two key challenges which mar existing face recognition systems: face recognition in scenarios of scarce data such as sketch to photo matching, and bias in automated facial analysis systems.

We begin by exploring the challenging problem of heterogeneous face recognition with scarce data. One such application is forensic sketch to digital face image matching, where a sketch drawn by a forensic artist is required to be matched against a database of high resolution face images. It is an important challenge with great social impact involving matching data from different domains. Improving sketch-face recognition can help law enforcement agencies perform a first-level filtration of the closest matches for the generated sketches, thus improving efficiency. In order to address this problem, we develop a transform learning based algorithm, *DeepTransformer*, which is feature agnostic in nature, and can be applied to existing features for enhancing the performance of a system. We extended this thread to other applications with scarce data such as skull to digital image matching and caricature to digital image matching for profile linking scenarios as well. DeepTransformer suffers from the challenge of limited feature-level discriminability and relatively large number of learnable parameters. In order to mitigate the above challenges, an effective and novel framework is proposed, termed as *Discriminative Shared Transform Learning* for cross-domain matching applications. A shared transform learns features in a common space for data belonging to different domains, and requires lesser number of parameters to be learned, thereby making it a suitable choice for scenarios with scarce data. The shared transform is learned while modeling the class variations, therefore effectively handling the increased inter-class similarity and high intra-class distance for cross-domain applications.

The later part of this dissertation focuses on studying another challenge which is affecting the current state-of-art automated facial analysis systems: biased performance with respect to specific sub-groups. Understanding and mitigating the effect of bias is of utmost importance, given the severe effects it can have in our society with rapid growth and deployment of AI based systems. We first perform in-depth analysis of deep learning based face recognition models for factors such as race and age. We observe the similarity in the behaviour of deep learning systems to human behavior as has been observed in several cognitive studies, in terms of the most discriminative regions learned by the model and used by humans, as well as the presence of the in-group effect. As the next step, this thesis presents mitigation strategies to de-bias existing models as well as learn fair models while training from scratch. A *filter drop* technique is presented, which is based on identifying filters responsible for learning the biasing/protected variable label. The technique involves dropping the filters and updating the model iteratively in order to perform unbiased classification.In order to eliminate the need for additional labels, a novel unbiased feature learning loss function termed as *Detox loss* is proposed. The proposed loss learns unbiased deep learning models to mitigate bias from existing networks, even with imbalanced data with respect to the protected attribute. It acts as an additional constraint which is a fairness constraint when training the model with the traditional classification loss. The Detox loss enforces that the learned features are distinguished based on the task label only, and not on the biasing-attribute. The results from the analysis and mitigation strategies can be extended for more generic applications as well, thus creating a positive impact on the society and the scientific community as a whole.

Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

List of Tables

Acronyms

CACD Cross Age Celebrity Dataset.

CMC Cumulative Match Characteristics.

CNN Convolutional Neural Network.

COTS Commercial Off-The-Shelf.

CSA Composite Sketch with Age Variations Dataset.

CUFS CUHK Face Sketch Dataset.

CUFSF CUHK Face Sketch FERET Dataset.

e-PRIP Extended PRIP Dataset.

FAR False Accept Rate.

GAR Genuine Accept Rate.

HOG Histogram of Oriented Gradients.

LBP Local Binary Patterns.

PRIP-VSGC PRIP Viewed Softtware Generated Composite Dataset.

RDF Random Decision Forest.

ResNet Residual Network.

ROC Receiver Operating Characteristic.

ROI Region of Interest.

SENet Squeeze and Excitation Network.

SVM Support Vector Machine.

Research Dissemination

Publications Related to the Dissertation

Journals and Selected Conferences:

1. **S. Nagpal**, M. Singh, R. Singh, and M. Vatsa, In-group Bias in Deep Learning based Face Recognition Models due to Ethnicity and Age, *IEEE Transactions on Technology and Society*, 2022 (*Minor Revision*)

2. **S. Nagpal**, M. Singh, R. Singh, and M. Vatsa, Detox Loss: Fairness Constraints for Learning with Imbalanced Data, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2022 *(Major Revision)*

3. **S. Nagpal**, M. Singh, R. Singh, and M. Vatsa, Discriminative Shared Transform Learning for Cross-Domain Matching, *Pattern Recognition*, vol.114: 107815, 2021 **(Impact Factor: 8.52)**

4. R. Singh, A. Agarwal, M. Singh, **S. Nagpal**, M. Vatsa, On Robustness of Face Recognition Algorithms Against Attacks and Bias, *AAAI Conference on Artificial Intelligence*, vol.34, no. 09, pp. 13583-13589, 2020

5. **S. Nagpal**, M. Singh, R. Singh, M. Vatsa, A. Noore, and A. Majumdar, Face Sketch Matching via Coupled Deep Transform Learning, *International Conference on Computer Vision*, pp. 5419-5428, 2017

6. **S. Nagpal**, M. Vatsa, and R. Singh, Sketch Recognition: What Lies Ahead? *Image and Vision Computing*, vol.55, pp. 9-13, 2016 **(Impact Factor: 3.86)**

Other Peer-Reviewed Conferences and Workshops:

7. M. Singh, **S. Nagpal**, D. Yadav, N. Kohli, P. Pandey, G. Prabhakaran, R. Singh, M. Vatsa, A. Noore, J. Brefczynski-Lewis, and H. Mahajan, Understanding Neural Responses to Face Verification of Cross-Domain Representations, *International Joint Conference on Neural Networks*, pp. 1-8, 2021

8. **S. Nagpal**, M. Singh, R. Singh, and M. Vatsa, Attribute Aware Filter-Drop for Bias-Invariant Classification, *IEEE Computer Vision and Pattern Recognition Workshop on Fair, Data Efficient And Trusted Computer Vision*,pp. 147-153, 2020

9. **S. Nagpal**, M. Singh, R. Singh, and M. Vatsa, Diversity Blocks for De-biasing Classification Models, *IEEE International Joint Conference on Biometrics*, pp. 1-9, 2020

10. M. Agarwal, S. Sinha, M. Singh, **S. Nagpal**, R. Singh, and M. Vatsa, Triplet Transform Learning For Automated Primate Face Recognition, *IEEE International Conference on Image Processing*, pp. 3462-3466, 2019

11. M. Singh, **S. Nagpal**, R. Singh, M. Vatsa, and A. Noore, Learning A Shared Transform Model for Skull to Digital Face Image Matching, *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pp. 1-7, 2018

12. **S. Nagpal**, M. Singh, A. Jain, R. Singh, M. Vatsa, and A. Noore, On Matching Skull to Digital Face Images: A Preliminary Approach, *IEEE/IAPR International Joint Conference on Biometrics*, pp. 813-819, 2017

13. T. Chugh, M. Singh, **S. Nagpal**, R. Singh, and M. Vatsa, Transfer Learning based Evolutionary Algorithm for Composite Face Sketch Recognition, *Computer Vision and Pattern Recognition Workshop on Biometrics*, pp. 619-627, 2017

Other Publications

Journals and Selected Conferences:

14. M. Singh, **S. Nagpal**, R. Singh, and M. Vatsa, DeriveNet for (Very) Low Resolution Image Classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2021.3088756, 2021

15. M. Singh, **S. Nagpal**, R. Singh, and M. Vatsa, Disguise Resilient Face Verification, *IEEE Transactions on Circuits and Systems for Video Technology*, doi: 10.1109/TCSVT.2021.3120772, 2021

16. M. Singh, **S. Nagpal**, R. Singh, and M. Vatsa, Dual Directed Capsule Network for Very Low Resolution Recognition, *IEEE/CVF International Conference on Computer Vision*, pp. 340-349, 2019

17. M. Singh, **S. Nagpal**, M. Vatsa, and R. Singh, Are You Eligible? Predicting Adulthood from Face Images via Class Specific Mean Autoencoder, *Pattern Recognition Letters*, vol.119, pp. 121-130, 2019 **(Impact Factor: 2.81)**

Other Peer-Reviewed Conferences, Workshops and Book Chapters:

18. M. Singh, **S. Nagpal**, M. Vatsa, and R. Singh, Enhancing Fine-Grained Classification for Low Resolution Images, *International Joint Conference on Neural Networks*, pp. 1-8, 2021

19. S. Gupta, N. Gupta, S. Ghosh, M. Singh, **S. Nagpal**, M. Vatsa, and R. Singh, FaceSurv: A Benchmark Video Dataset for Face Detection and Recognition Across Spectra and Resolutions, *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1-7, 2019

20. I. Kalra, M. Singh, **S. Nagpal**, R. Singh, M. Vatsa, and P.B. Sujit, DroneSURF: Benchmark Dataset for Drone-based Face Recognition, *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1-7, 2019

21. **S. Nagpal**, M. Singh, M. Vatsa, R. Singh, and A. Noore, Expression Classification in Children Using Mean Supervised Deep Boltzmann Machine, *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop on Analysis and Modeling of Faces and Gestures*, pp. 236-245, 2019

22. M. Singh, **S. Nagpal**, R. Singh, M. Vatsa, and A. Majumdar, Identity Aware Synthesis for Cross Resolution Face Recognition, *International Conference on Computer Vision and Pattern Recognition - Workshop on Biometrics*, pp. 592-59209, 2018

23. M. Singh, **S. Nagpal**, M. Vatsa, R. Singh, A. Noore, and A. Majumdar, Gender and Ethnicity Classification of Iris Images using Deep Class Encoder, *IEEE/IAPR International Joint Conference on Biometrics*, pp. 666-673, 2017 (**Received Best Poster Award**)

24. M. Singh, **S. Nagpal**, R. Singh, and M. Vatsa, Class Representative Autoencoder for Low Resolution Multi-Spectral Gender Classification, *International Joint Conference on Neural Networks*, pp. 1026-1033, 2017

25. D. Yadav, N. Kohli, **S. Nagpal**, M. Singh, P. Pandey, M. Vatsa, R. Singh, and A. Noore, Region-specific fMRI Dictionary for Decoding Face Verification in Humans, *International Joint Conference on Neural Networks*, pp. 3814-3821, 2017

26. **S. Nagpal**, M. Singh, M. Vatsa, and R. Singh, Deep Learning: Fundamentals and Beyond, *Deep Learning in Biometrics*, pp. 1-31. CRC Press, 2018

27. M. Singh, **S. Nagpal**, N. Gupta, S. Gupta, S. Ghosh, R. Singh, and M. Vatsa, Cross-Spectral Cross-Resolution Video Database for Face Recognition, *IEEE International Conference on Biometrics: Theory Applications and Systems*, pp. 1-7, 2016

Chapter 1

Introduction

Humans have the innate ability to learn, analyse, memorize, and perform cognitive functioning. We have been performing simple tasks such as counting to complex tasks such as identifying people based on their voice or face alike since time unknown. Traces of using numbers, calculations, and reasoning have been found in different parts of the world as far back as the Sumerians and Egyptians before 2000 BC. Human beings performed most calculations using bones, pebbles, counting boards, and the abacus. The beginning of the 17^{th} century saw the development of more complex arithmetic tools such as logarithms, slide rulers, and military compass. A few decades later, this was followed by the introduction of mechanical calculators in the Renaissance period. With technological advancements, came new methods to do the same task. One of the early known transitions to an automated device was the transformation from a mechanical calculator to an electronic calculator with the introduction of the mainframe, which began in the mid 20^{th} century. Within no time, we saw an array of new devices and computers being introduced, and today one of the most powerful calculators is actually a super computer capable of performing calculations in one second that a human would have taken six billion years [23]!

During the past century, the human dependence and use of devices in day-to-day tasks has grown exponentially. This was primarily fostered by the digital revolution during the same period. It is interesting to note that even though the digital transformation era was a long time coming, the transition and adaptation has been done overnight with new tools and technologies being developed at lightning speed.

As can be seen in Figure 1-1, access control and security has been a key aspect of our lives,

be it for kingdoms and homes to countries and restricted spaces. A primary method of doing so has been person identification. In the early days, this task was done by humans who would be responsible for securing a set up and would manually verify the identity before allowing access to any person. Since then, with evolution and advancement, locks and keys were introduced, followed by passcodes and password based authentication systems in controlled environments. However, in larger areas like airports, the identification was still being done by humans.

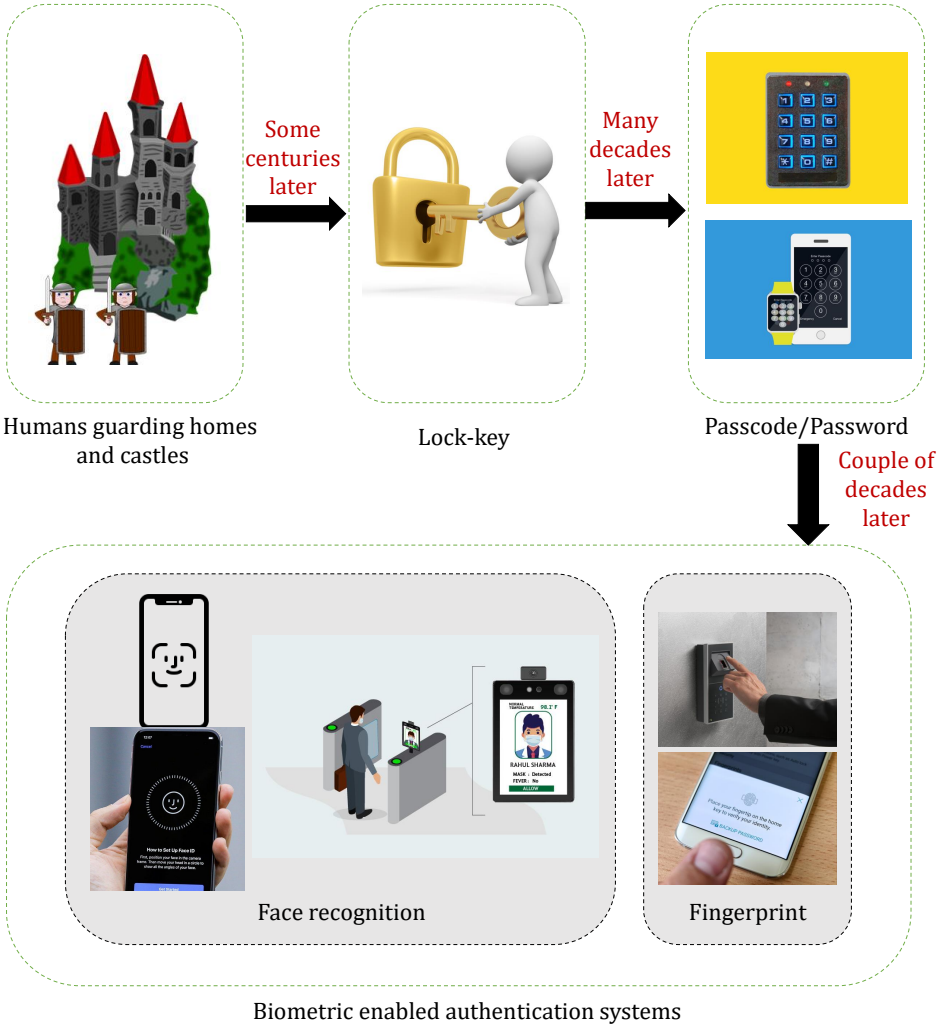


Figure 1-1: Evolution of tools and processes for authentication over the years. Houses and kingdoms were manned and humans would be contolling access to their spaces in the early years. With the industrial revolution came lock and key which were used for securing access. As we stepped into an era of technological advancement, this was followed by passcode and password based methods. The overnight digital trasnformation and advent of machine learning has introduced more secure biometric based authentication systems like fingerprint and more recently, non intrusive and seamless methods based on face recognition.

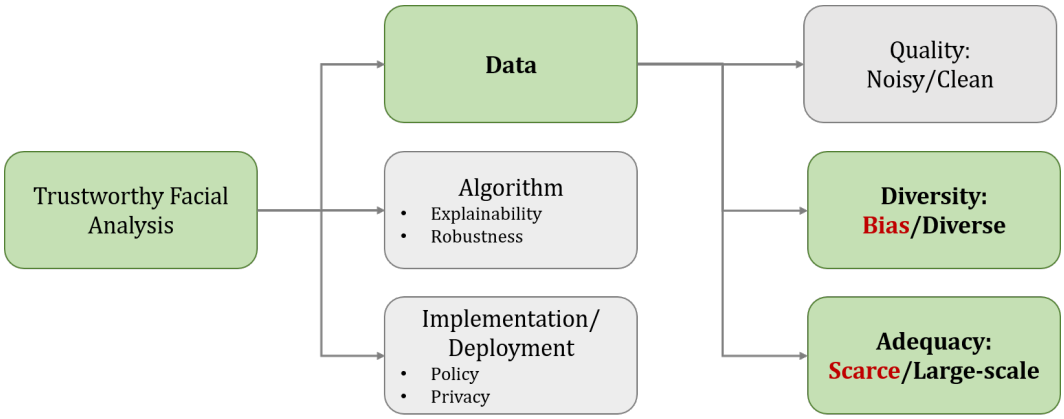


Figure 1-2: Trustworthy facial analysis relies on three components: data, algorithm, and deployment. This dissertation focuses on data centric challenges, specifically developing facial analysis models with scarce and biased data.

In the recent times, with the advent of machine learning algorithms and overnight digital transformation, automated models are being used for various tasks such as object classification for image tagging, resume shortlisting, credit review for loan applicants, caption generation for a given image, and spam versus ham classification of emails/messages. Automated models have also been developed and deployed to perform biometric based authentication. Fingerprints were the first candidate biometric to be used given their high accuracy and ease of use. They were used with the users' cooperation and knowledge. With further development, adaptation of digital interfaces and IoT in our day-to-day lives, as we move towards seamless integration of our devices and homes, face recognition is being used as a non-intrusive method to secure our spaces and offer access control [15].

Face recognition has widespread application today beyond physical and electronic access control. It is used from social media tagging and profile linking to facial analysis for targeted advertisements, protect law enforcement, find missing persons, aid visually challenged individuals, attendance, etc. It is a technology which has been riding on the digital transformation wave and has widely been accepted and adapted due to its applicability and ease of use. However, it is interesting to note that given the quick turnaround time in demand to deployment of technology, we focused on developing model-centric algorithms which work primarily in constrained setups. Further, limited attention was focused on developing trustworthy AI systems, which are of utmost importance. Figure 1-2 presents the different components of a trustworthy facial analysis model,

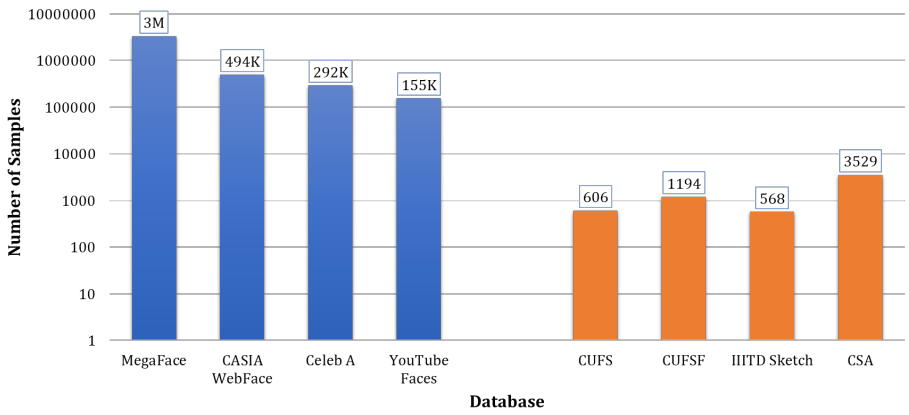


Figure 1-3: Graph demonstrating the number of images in popular face recognition (blue) and sketch face recognition (orange) datasets. The face recognition datasets contain 100k+ images, whereas popular sketch datasets contain at most 3500 images, resulting in the challenge of scarce and limited data.

where dedicated attention is required on the data, algorithm modeling, and deployment stages, respectively. While initial research in face recognition focused heavily on curating datasets and developing models under varying *physical covariates* such as pose, illumination, or expression [61, 85, 159], recent needs demand dedicated attention on the data and model characteristics to ensure trustworthiness. To this effect, this dissertation focuses on the data stage of building reliable models, which has large amount of impact both in terms of technology and society. Thus, this dissertation focuses on two key aspects of face recognition: with (i) scarce and (ii) biased data, and aims to understand and present data centric algorithms. The remainder of this chapter discusses the two aspects in detail followed by the research contributions of the dissertation.

1.1 Face Recognition with Scarce Data

Over the past decade, there has been an exponential improvement in performance achieved by algorithms for face recognition and analysis tasks [29, 110, 112, 145, 178]. Most recent approaches are deep learning based approaches and have been trained on large scale datasets. Some of the commonly used datasets are Casia WebFace [200], YouTube Faces [190], Mega Face [78], CelebA [102], etc. and each of them contain at least 100k images. Figure 1-3 showcases the number of images in some commonly used face recognition datasets (in blue). However, real world data contains unconstrained images and is often captured from different sources. This often leads to data

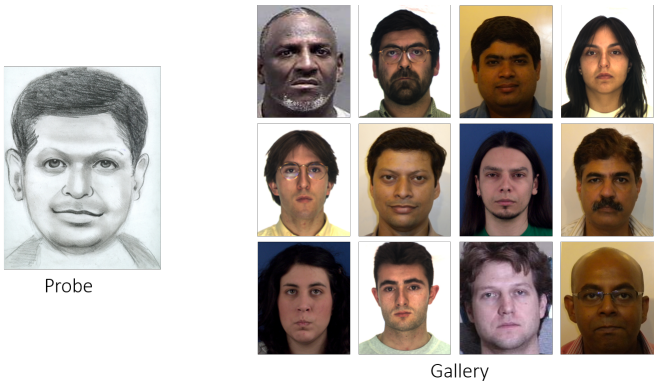


Figure 1-4: Sketch recognition refers to the task of finding the best match from the set of gallery images for a given query image. Images are taken from AR dataset [109], MMI dataset [131], IIITD Sketch dataset [18], CMU MultiPIE dataset [159], MORPH dataset[151].

differing in information content as well, introducing the challenge of heterogenity or heterogenous matching.

Heterogeneous face recognition refers to the task of matching face images belonging to different domains. Specifically, given data from two domains - domain 'A' and domain 'B', the query image belongs to domain 'A', the task is to find the best match from the given set of gallery images belonging to domain 'B'. One such use-case in the real world setup is that of automated sketch face recognition. As can be seen in Figure 1-4 sketch face recognition refers to the task of matching data pertaining to two different domains - sketch and digital image. The problem statement can be defined as: Given a query sketch image, the aim is to find the best match from the pool of images, also known as the gallery, which consists of digital images. Adding to the challenge of heterogenity, such datasets are often small in size and contain limited data. Figure 1-3 also shows the number of images present in some of the largest sketch face recognition datasets. We see that for tasks such as sketch face recognition, even the popular datasets contain limited data samples introducing the challenge of working with scarce or limited images.

Automated sketch to digital image matching has applicability in several day to day scenarios such as forensic sketch matching in law enforcement scenarios, or profile linking using caricature face images on social media, and beyond faces in cases such as similar object image retrieval. Some real world examples [156] are as follows:

- Sketch portraits are used widely by law enforcement agencies. In India, in 2016 alone, the State Crime Records Bureau generated at least 350 portraits of which at least 150 helped

police solve the cases and find the suspect.

- In another case in Sriperumbudur, wherein the face of a person was severely crushed, a sketch was generated based on the face markers and the image was widely circulated, which led to the identification of the subject.

- Automated sketch/caricature face recognition can also enable labelling of faces in digital media and result in better search results.

Despite the widespread applicability, there are several reasons which make sketch face recognition different and more challenging than general face recognition. The challenges associated with sketch to digital image matching are summarized below:

- **Domain gap:** Given that the images belong to different domains, the type of information content in both the domains varies greatly. Sketch images are edge information rich, lack texture information, and primarily contain high frequency information. Whereas, digital images are texture rich and contain high frequency as well as low frequency information.

- **Limited availability of labeled/mated data:** Sketch to digital face matching requires mated pairs or corresponding image pairs which are often difficult to obtain.

- **Difference of perspective:** Since there is no single technique for sketch generation, it often results in varying artistic or software styles, along with the variation in perspective and interpretation leading to artistic or software bias to be introduced at the time of creating the sketch.

- **Forgetting process:** This challenge is majorly specific to the sketches created in surveillance or crime scene scenarios (also known as forensic sketches) which are generated based on the memory of humans and the description provided by them either to an artist or a software operator. Since the desired sketches are being created based on the memory of the witness, it is important to note that due to the behaviour and functioning of the human brain, there will be some level of details which will be missed or forged due to the forgetting process of the human brain.

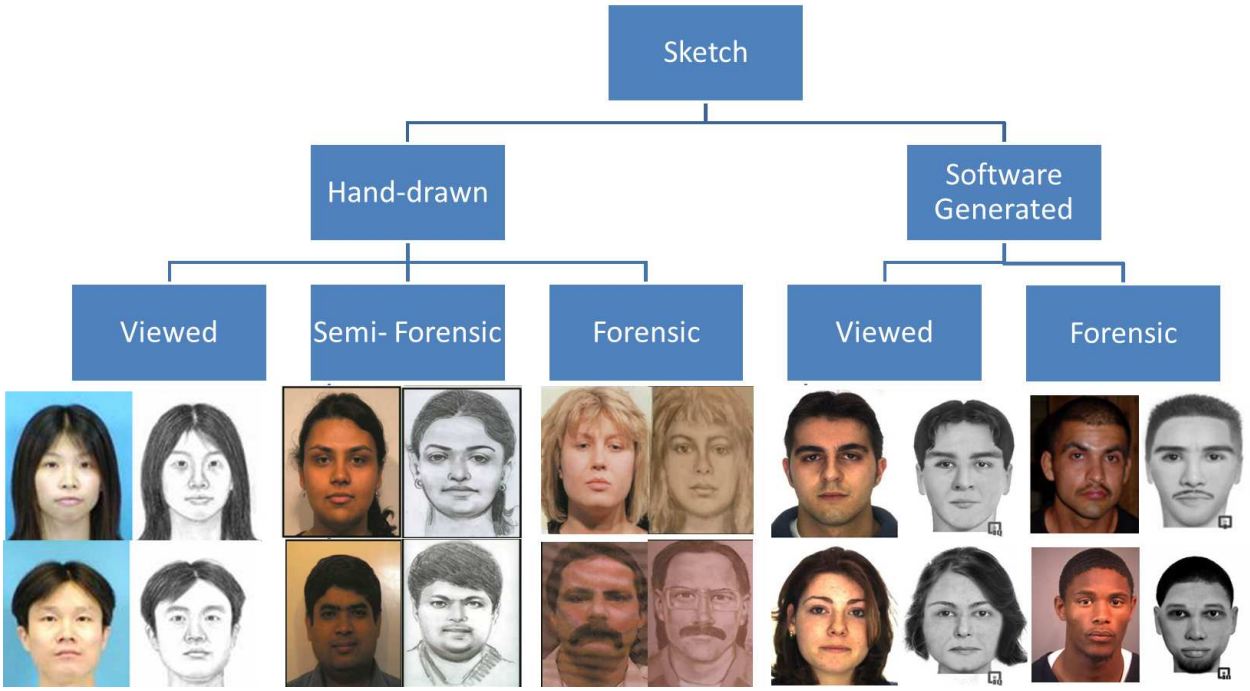


Figure 1-5: Illustrating samples of different types of sketches: hand-drawn and software generated. Images are taken from [185], [18], and world wide web.

Facial sketches can broadly be divided into two categories: hand-drawn and software-generated composites. Based on the description of an eye-witness, hand-drawn sketches are drawn by an expert sketch artist, whereas composite sketches are created via specialized software by trained technicians. As shown in Figure 1-5, sketch generation process is further divided into three categories:

- Viewed sketch: when the artist/technician draws the sketch by observing the person or an image of the person,
- Semi-forensic sketch: when the artist draws the sketch based on his/her memory of the person/image, and
- Forensic sketches: a forensic artist draws the sketches based on the description provided by an eye-witness.

Initial research on sketch face recognition began with generating datasets of viewed sketches. This was superseded by the introduction of semi-forensic sketches to incorporate the time lag be-

tween the artist looking at the digital image and creating the sketch. Viewed and semi-viewed datasets acted as the stepping stone for initiating research in this domain for the covariate of sketches. Given that the primary challenge introduced in these sketches was that of domain gap, the approaches in literature can broadly be classified into the following:

- Learning domain invariant features: In these approaches, the aim is to learn features which do not depend upon the domain of the image, i.e. can be used for matching irrespective of the domain [18, 116, 172].
- Synthesis based: In these approaches, data is often transformed from one domain to the other and then matching is performed once the data is in the same domain. [117, 183].
- Transform data to a common space: Data can also be transformed to a common space. In such an approach, data from both domains is transformed to a common space between the two domains in order to prevent loss of information from either of the domains [97].

CUFS [185], CUSFS [209], and IIITD Sketch [18] databases are some of the popular viewed and semi-forensics hand-drawn sketch databases. While these sketches are important for research, forensic sketches are of paramount interest and are used in real-world law enforcement applications. Except few samples that are made available (via their books) by Lois Gibson, Karen Taylor and other sources on the World Wide Web, there is a scarcity of hand-drawn forensic sketch databases for research. Using sketch generation softwares for research is still in its nascent stage and only PRIP [64] and e-PRIP databases [117] are available for research.

However, within a short span of time, researchers were able to obtain high accuracies on the simulated datasets, and it was soon realised that these datasets are unable to capture the real essence of the problem of sketch to digital matching and are not very representative of the real scenario. Additionally, none of these simulated datasets also captured or represented other use cases such as caricature recognition, or sketch to sketch matching with different types of sketches for profile linking use cases. Thus, it is of utmost importance to look at the real world data and develop approaches specific to the challenges at hand for the above mentioned tasks. This research primarily focuses on the real world forensic data and on developing a learning paradigm which enables us to develop algorithms to perform classification on real sketch data as well as work on various other relevant use cases such as sketch to sketch matching and caricature face recognition.

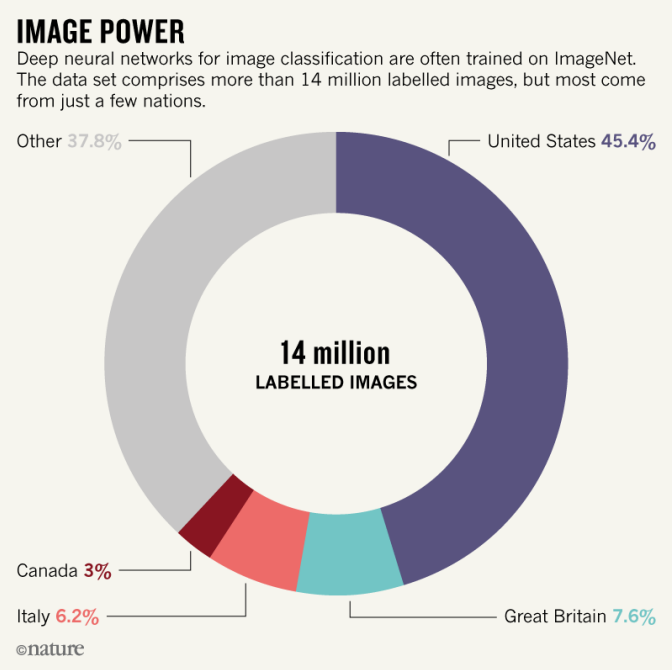


Figure 1-6: Location-wise distribution of the samples in the commonly used ImageNet dataset [90] [Source: https://www.nature.com/articles/d41586-018-05707-8].

1.2 Face Recognition with Biased Data

While scarce data is one of the prevelent challenges for developing efficient AI models, recently, the machine learning community has been marred with the challenge of bias and fairness in AI systems [24, 66, 88, 119, 143, 168]. Researchers have presented several case studies demonstrating bias in various applications and problems such as face recognition, attribute prediction, activity recognition, and automated caption generation. In one such instance, a commercially deployed face based gender classification software has shown to be more accurate for lighter skinned males in comparison to darker skinned females [24]. In another incident, an image based skin cancer detection algorithm was developed, only to later realize that the data used for training contained less than 5% dark skinned woman, thus the algorithm was not tested on dark skinned women [216].

Today, machine learning systems are omnipresent in our lives, therefore it is essential to develop unbiased models, which do not present unfair outcomes. The predictions of these learned models should not be based on or biased against a particular sensitive attribute, thus resulting in bias-free outcomes. Unfair outcomes can have severe implications depending on the task of the machine learning models. For example, an automated recruitment tool should make hiring deci-

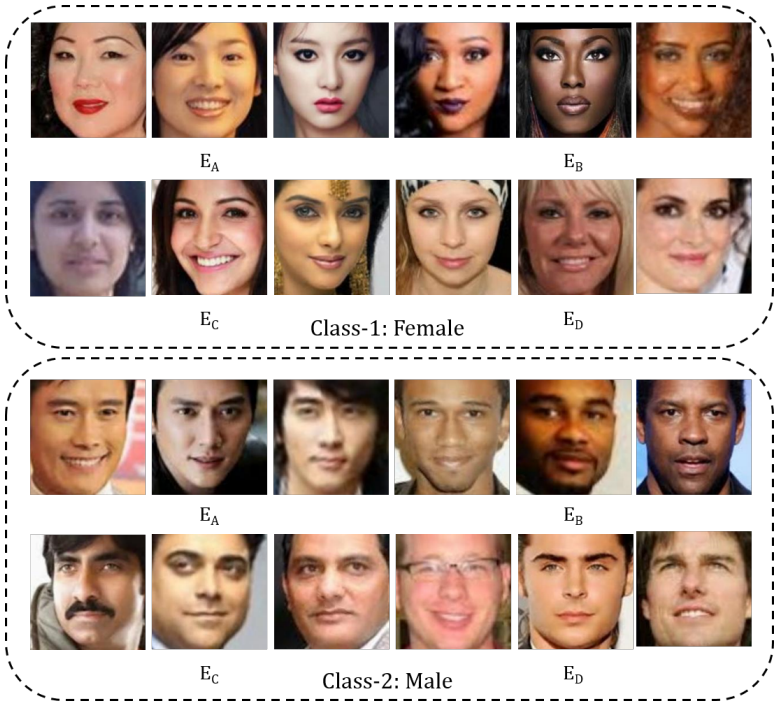


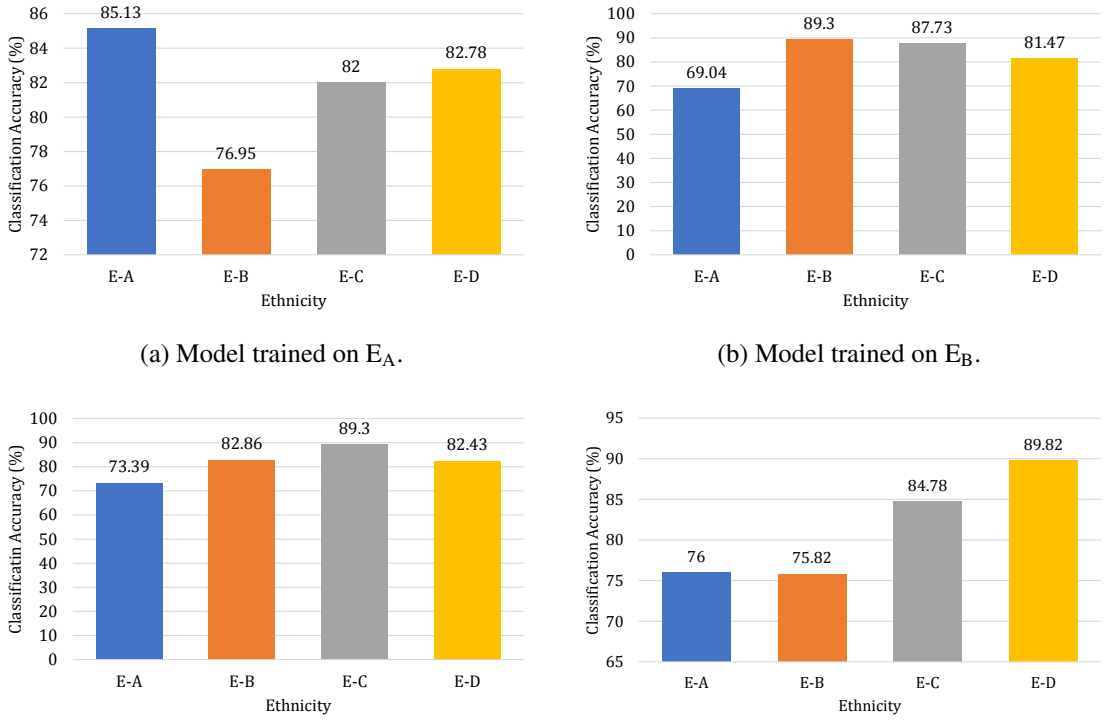
Figure 1-7: Sample images from the UTKFace dataset [211] (four ethnicities) used for gender prediction. The top two rows and the bottom two rows present female and male face images, respectively.

sions based solely on the professional qualifications, without any inherent bias due to some other factor such as gender or age-group.

Existing studies have shown that the presence of bias can either be attributed to (i) the tainted examples which promote human bias against a particular sub-group, or (ii) skewed training samples which lead to imbalanced data with respect to a particular sub-group (Figure 1-6) [14]. In computer vision, bias has usually been observed due to skewed training datasets. For instance, for a facial analysis model, the training samples might not be balanced with respect to an attribute such as gender or ethnicity.

In order to observe the existence of bias and further understand the biased behavior of models, we simulate a toy example of facial analysis based gender classification, in a restricted experimental setting. The simulation is performed using the UTKFace dataset [211] for gender prediction with multiple ethnicities (each ethnicity is treated as a sub-group) [1]. The UTKFace dataset contains over 20,000 face images from different ethnicities and age-groups, containing an ethnicity label

[1]This toy example has been published in the IEEE International Joint Conference on Biometrics, 2020, and all the corresponding figures have been taken from the published manuscript [121].



(a) Model trained on E_A.

(b) Model trained on E_B.

(c) Model trained on E_C.

(d) Model trained on E_D.

Figure 1-8: Classification accuracy obtained on face images of different ethnicities (E_A, E_B, E_C, and E_D), when the model is trained on a subset of face images only. All the images have been taken from the same dataset (UTKFace dataset). The figure has been taken from the published manuscript [121].

and a gender label (male/female) for each image. Therefore, we have used the UTKFace dataset for analyzing gender classification models with respect to ethnicity bias. Training and testing partitions are generated for four ethnicities: E_A, E_B, E_C, and E_D[2] (Figure 1-7). The number of images in the train and test sets were kept consistent across ethnicities (randomly chosen 2000 and 1150, respectively), while ensuring equal number of images for each class (male/female). Experiments have been performed using the LightCNN-9 [192] architecture, consisting of nine convolutional layers and ten Max-Feature-Map layers.

Within-dataset Cross-Ethnicity Bias: In order to understand and establish the cross-ethnicity bias, a LightCNN-9 model is trained for the task of gender prediction, using data pertaining to a

[2]Since the aim is to study cross-subgroup effect, and not draw ethnicity-specific analysis, the ethnicity information has been anonymized. The ethnicities as defined in the UTKFace database are - E_A: Asian, E_B: Black, E_C: Indian, E_D: White.

single ethnicity only. Evaluation is performed using the model trained on face images of individuals from different ethnicities. Figure 1-8 presents the gender prediction accuracies obtained using different models (trained on a single ethnicity) on the test set of E_A, E_B, E_C, and E_D. It is important to note that the test set remains consistent across all the experiments. These results showcase that despite drawing from the same class as in the source, the performance of different ethnicities varies substantially depending upon the training data used. The best performance on each ethnicity is obtained by using the model trained on that ethnicity's data. For example, 85.13% is obtained for E_A using the model trained on E_A (Figure 1-8(a)), which is at least 9% higher than the performance obtained by any other model trained on other ethnicities. In cross-ethnicity training-testing, when samples from E_A are evaluated using the model trained on E_B, gender prediction accuracy of 69.04% is obtained (Figure 1-8(b)), demonstrating a variation of almost 15%. Similar accuracy trends are observed for the other ethnicities as well, suggesting bias due to ethnicity variations, and the need for robust de-biasing techniques.

Effect of Fine-tuning: As demonstrated above, the problem of bias exists and goes beyond the phenomenon of 'dataset bias', wherein models trained on a particular dataset perform poorly on another dataset. One of the most common techniques for handling such mis-classifications in deep learning is *fine-tuning*. A pre-trained model can be fine-tuned using a subset of data from the target data distribution to obtain improved performance. In order to evaluate the effectiveness of the existing fine-tuning methodology, analysis is performed on the same toy example. From Figure 1-8 it can be observed that the maximum variation in accuracy is observed between E_A and E_B, that is, the model trained on E_A presents an accuracy variation of almost 9% between E_A and E_B (Figure 1-8(a)), while the model trained on E_B presents an accuracy variation of 20% between E_A and E_B (Figure 1-8(b)). Therefore, in order to understand the effectiveness of fine-tuning as a possible solution for de-biasing, the gender prediction model trained on E_A is fine-tuned with the training data belonging to E_B. Figure 1-9 presents the classification performance of the base model trained on E_A only, as well as the performance obtained after fine-tuning with E_B. The base model presented an accuracy of 85.13% and 76.95% on E_A and E_B, respectively. After fine-tuning, the accuracy on E_A dropped to 74.50% (over 12% reduction), while the performance on E_B improved drastically to 88.69% (over 11% improvement). While fine-tuning a pre-trained model enhances

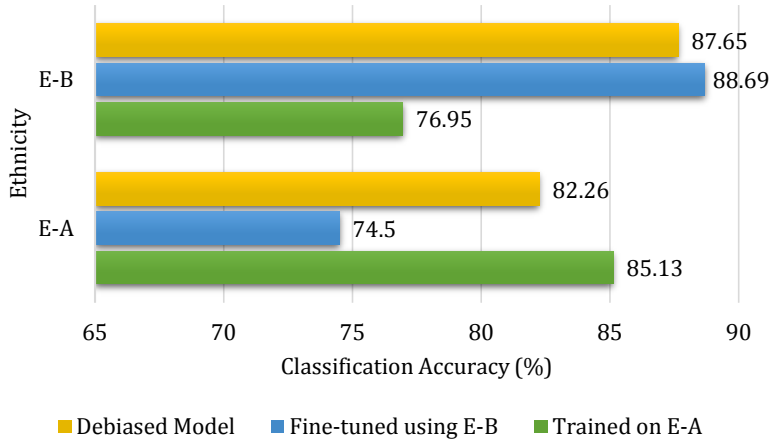


Figure 1-9: Demonstrating the effect of fine-tuning the model trained on E_A with face images of E_B, and the effect of including the proposed diversity block for de-biasing. The figure has been taken from the published manuscript [121].

the performance on the E_B face images, it led to a severe reduction in the classification performance on E_A, which is not a desirable phenomenon for a de-biased model. Thus, fine-tuning appears to result in the model *forgetting* the previously learned information [60].

The classification error propagated in the biased models highlights a deep rooted challenge which mars the AI community. The classification error is not a random mis-classification, but bears a severe connotation with respect to *fairness* [166]. Most constitutions across the globe provide it's citizens with the Right To Equality, which ensures no individual is discriminated against based on their gender, caste, religion, race or birthplace. Until now, the right to equality is expected to be preserved, whenever humans are involved in the decision making process. However, there is no similar enforced mandate for ensuring unbiased decisions of automated models. In case a model favors a particular subgroup based on the gender, color, caste, or birthplace, it leads to *unfair predictions*, which is a direct violation of the Right to Equality. To this effect, research efforts have been made in various directions to understand bias and debias models. Initially, research was focused on studying the existence of bias in Automated models. Multiple studies have highlighted the existence of bias in automated systems. De Vries *et al.* [41] demonstrated the poor performance of publicly available object classification cloud services on the data belonging to different income groups and geographical locations. Robinson *et al.* [153] proposed the Balanced Faces in the Wild dataset containing data balanced across gender and ethnicity. The authors reported

the accuracy variations across sub-groups and proposed learning sub-group specific thresholds. Buolamwini and Gebru [24] demonstrated the poor performance of commercially deployed facial analysis software on a subgroup of dark skinned females in comparison to lighter skinned males. The authors performed an in-depth analysis by utilizing phenotypic labeling of data using the six point Fitzpatrick classification system to showcase the biased behaviour of the software. Based on several studies, it was understood that there is a need to estimate bias, evaluate models for bias and also learn unbiased models. Manjunatha *et al.* [108] presented a bias identification technique for visual question answering models. Traditionally, a machine learning classification pipeline consists of three stages: (i) input, (ii) feature extraction, and (iii) classification. Researchers have presented techniques to learn unbiased representations by either presenting a pre-processing technique on the input data, or by modifying the feature extraction stage, both of which require the model to be trained from scratch or re-trained. For instance, Amini *et al.* [11] presented a debiasing technique for face detection algorithms. The authors learned the latent structure in the training data with respect to the ethnicity and gender of the subject via variational autoencoders, which was utilized to re-weight samples. Ryu *et al.* [155] proposed Inclusive FaceNet model, which utilized transfer learning to learn attribute prediction models for various subgroups across gender and ethnicity. Recently, techniques have also been proposed for learning frameworks to perform unbiased classification. Multi-task Convolutional Neural Network (MTCNN), presented by Das *et al.* [39], jointly learns to predict the gender, ethnicity, and age from the input. Joint learning results in improved learning across sub-groups which reduces the biased behavior of the model towards a particular sub-group. A joint learning and unlearning framework [10] has also been proposed for eliminating bias from CNN models for age, gender, race, and pose classification from face images. Additionally, Hendricks *et al.* [66] investigated the gender specific words at the time of caption generation from images, and proposed an Equalizer model which ensures equal probability for the gender when gender information is occluded.

Despite several research efforts to establish and evaluate the existence of bias, limited or no efforts have been made to understand how and where the bias exists, particularly in facial analysis models. Thus, this dissertation studies bias in facial analysis models and evaluates the effect of in-group effect in deep learning models. Based on the findings, we understand where the bias is and what causes it. The second main focus of this research in terms of biased data is to present

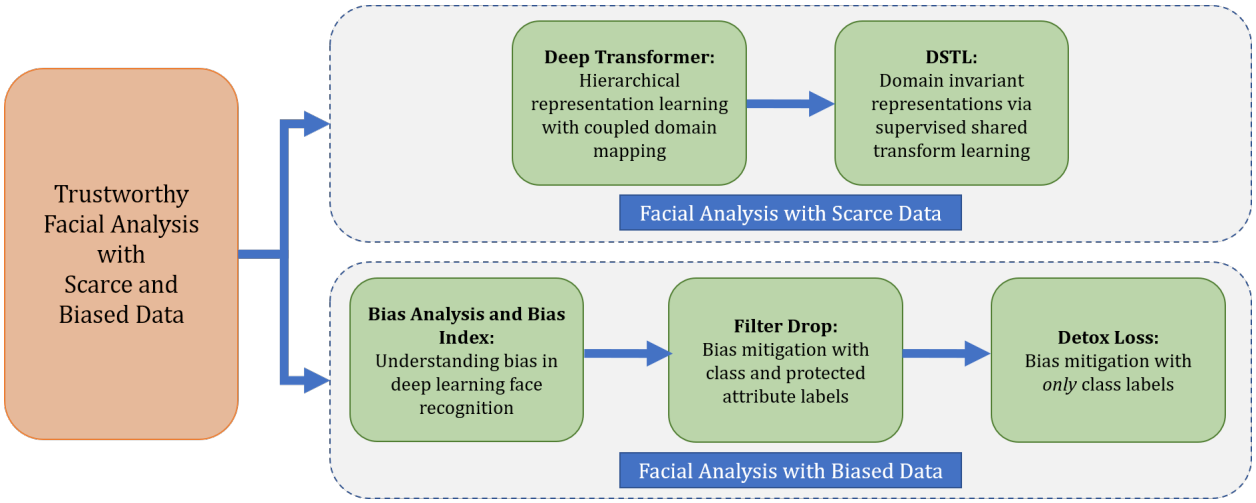


Figure 1-10: Key contributions of this thesis focusing on trustworthy facial analysis with scarce and biased data.

algorithms to learn unbiased models for facial analysis tasks.

1.3 Research Contributions

With the need for secure systems, ease of monitoring, authentication, attribute analysis and widespread applicability of digital media, face recognition and analysis systems have become omnipresent. Research had primarily focused on face recognition and analysis in constrained setups. As mentioned previously, scarce and biased data has been generally overlooked while developing model-centric approaches. To bridge this gap and move towards developing trustworthy facial analysis systems, the core novelty of this dissertation revolves around exploring and developing algorithms for scarce and biased data for facial analysis and recognition (Figure 1-10). Often, in real world setups, facial analysis is marred with the challenge of limited availability of data in scenarios of heterogenous setups such as sketch to digital image matching, or thermal image to digital image matching. In this thesis, we focus on sketch to digital image matching as the task for scarce data and develop algorithms for the same. Representation learning based algorithms are presented for sketch face recognition to demonstrate the efficacy of the proposed approaches with scarce and limited data. Along with scarce data, another key challenge overlooked in the era of deep learning during model development is of biased data. This has implications beyond performance of AI models and can lead to unfair outcomes. This thesis delves deeper to explore automated algorithms with biased

data by understanding how deep learning based approaches suffer from the challenge of bias and how it contrasts with human bias. Two novel algorithms are also presented to mitigate the effect of biased data at the time of training. The key contributions of this dissertation are as follows:

- **DeepTransformer for Face Recognition for Scarce Data:** As discussed previously, several tasks in face recognition are marred by the availability of limited data. Face sketch to digital image matching is one such task. It is an important challenge of face recognition that involves matching across different domains. This dissertation presents a novel transform learning based approach termed as DeepTransformer, which learns a transformation and mapping function between the features of two domains. The proposed formulation is independent of the input information and can be applied with any existing learned or handcrafted feature. Since the mapping function is directional in nature, we propose two variants of DeepTransformer: (i) semi-coupled and (ii) symmetrically-coupled deep transform learning. The performance of the proposed models is evaluated on a novel application of sketch-to-sketch matching, along with sketch-to-digital photo matching. Experimental results demonstrate the robustness of the proposed models.

- **Discriminative Shared Transform Learning (DSTL) for Scarce Data:** In order to address the challenges elaborated above, this research proposes a novel *Discriminative Shared Transform Learning (DSTL)* algorithm for sketch to digital image matching. DSTL learns a shared transform for data belonging to the two domains, while modeling the class variations, resulting in discriminative feature learning. Two models have been presented under the proposed DSTL algorithm: (i) Contractive Model (C-Model) and (ii) Divergent Model (D-Model), which have been formulated with different supervision constraints. Experimental analysis on seven datasets for three case studies of sketch to digital image matching demonstrate the efficacy of the proposed approach, highlighting the importance of each component, its input-agnostic behavior, and improved matching performance.

- **Understanding In-group Bias in Deep Learning based Face Recognition Models and Predicting Level of Bias Via novel Bias Index:** Existing literature shows that humans exhibit biased behaviour, wherein people tend to favor other individuals who exist in similar groups as them, and termed it as *in-group* bias [22, 93, 157]. The groups could be formed

on the basis of ethnicity, age, or even a favorite sports team. Taking cues from these findings, this dissertation inspects if deep learning networks also mimic this human behaviour, and are affected by *in-group* and *out-group* bias. The behavior of face recognition models is evaluated to understand if, similar to humans, they encode *group-specific* features for face recognition, along with *where* bias is encoded in face recognition models. Specifically, analysis has been performed for two use-cases of bias due to ethnicity and age in face recognition models. Extensive experiments involving multiple datasets are performed using a collection of known and our own trained deep networks brings forth several interesting insights suggesting behavior similar to humans. The experimental evaluation showcases that deep learning models focus on different facial regions for different ethnicity or age. Large variation in face verification performance is also observed across different sub-groups for both known and our own trained deep networks. For example, a variation of almost 24% is observed between different age-groups for face verification using large-scale pre-trained models. To the best of our knowledge, this is the first of its kind research which inspects deep networks for exhibiting in-group effect to understand bias in deep learning systems. Based on the observations, this dissertation also presents a novel *bias index* for evaluating a trained model's *level of bias* by analyzing the learned regions of interest from class activation maps. Understanding of the state-of-the-art deep learning models and the proposed bias index is essential to address the challenge of bias in AI, and develop fairer algorithms.

- **Attribute Aware Filter-Drop for Unbiased Classification:** The widespread applicability of deep learning based algorithms demands dedicated attention towards ensuring unbiased behavior. Biased feature learning (for or against a particular sub-group) might often result in unfair predictions. In this dissertation, a novel *Filter-Drop* algorithm for learning unbiased representations is presented. The proposed technique focuses on learning the features useful for predicting the biasing attribute (or the sensitive attribute), followed by their elimination while performing the primary classification task. To this effect, a multi-task network is trained, which prevents the features capturing the attribute variations from being used for the primary classification task. The efficacy of the proposed Filter-Drop technique is demonstrated on two facial analysis datasets: UTKFace dataset and FairFace dataset. The proposed

technique achieves similar performance across different ethnicity groups while training with highly skewed training data as well.

- **Detox Loss: Fairness Constraints for Learning with Imbalanced/ Biased Data:** Building upon the previous contributions and in order to eliminate the need of an additional label (biasing attribute), this dissertation proposes *Detox loss*, a novel feature learning loss function for learning unbiased models. The proposed loss can be used to: (i) learn fairer deep learning classifiers, and (ii) mitigate bias from existing pre-trained networks, especially in the challenging constraint of imbalanced training data with respect to a protected attribute. Conceptually, the Detox loss enforces that the learned features are distinguished based on the task label only, while eliminating any distinction based on the biasing-attribute. This is achieved by incorporating three fairness constraints while training with the traditional classification loss: (i) proposed *Exclusion loss*, (ii) proposed *Inclusion loss*, and (iii) proposed *Feature-distillation loss*. The efficacy of the Detox loss is demonstrated on two facial analysis tasks: (i) age-group prediction and (ii) gender prediction, under the protected attribute of ethnicity for learning fairer models and de-biasing existing models, with varying imbalanced training data distributions. Across two different protocols, three setups, and two tasks, the Detox loss obtains state of the art performance without the need of multi-labeled training data. For example, on the challenging Pilot Parliaments Benchmark (PPB) dataset, the Detox loss obtains a balanced accuracy and F_1 score of 96.1% and 96.6%, as compared to the state-of-the-art performance of 94.1% and 93.6%, respectively.

Chapter 2

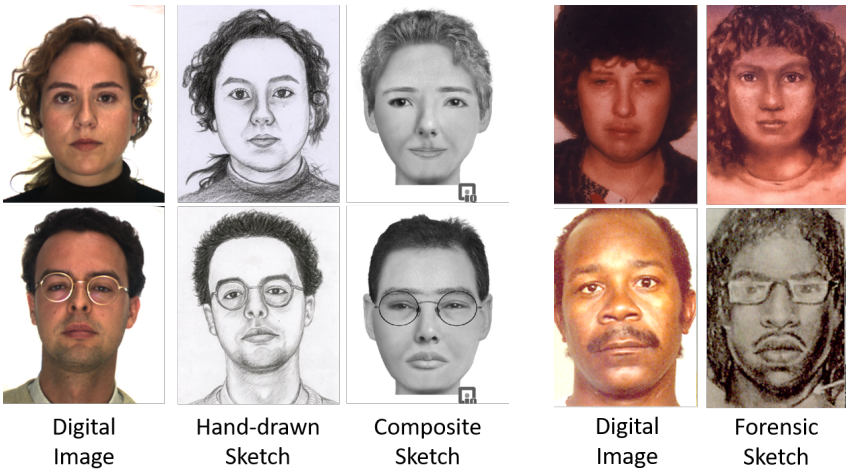
Face Sketch Matching via Coupled Deep Transform Learning

2.1 Introduction

Face recognition systems have been evolving over the past few decades, particularly with the availability of large scale databases and access to sophisticated hardware. Large scale face recognition challenges such as MegaFace [78] and Janus [85] further provide opportunities for bridging the gap between unconstrained and constrained face recognition. However, the availability of new devices and applications continuously open new challenges. One such challenging application is matching sketches with digital face photos. In criminal investigations, eyewitnesses provide a first hand account of the event, along with a description of the appearance of the suspect based on their memory. A sketch artist interviews the eyewitness of a particular case and a sketch image of the suspect is created. Such a sketch drawn by an artist is termed as a *hand-drawn sketch*. To eliminate the inter-artist variations and automate the process of sketch generation, law enforcement agencies have started using software generated *composite sketches*. In such cases, the eyewitness is interviewed by an officer and a sketch is created using the drag-and-drop features available in sketch generation tools such as FACES [2], evoFIT [1] and IdentiKit [3]. As shown in Figure 2-1(a), the information content in the two domains/modalities (sketch and digital image) vary significantly. The digital image is an information-rich representation whereas, the sketch image contains only the edge information and lacks texture details. Figure 2-1(b) shows real world examples of foren-

Table 2.1: A brief literature review of sketch-to-photo matching problem.

Sketch	**Authors (Year)**	**Feature Extraction**	**Classification**
Hand-drawn	Bhatt *et al.* [18]	Proposed MCWLD	Memetically optimized chi-squared distance
	Khan *et al.* [81]	Facial Self Similarity descriptor	Nearest neighbor classifier
	Mignon *et al.*[114]	Proposed Cross modal metric learning (CMML)	
	Klare *et al.* [83]	MLBP, SIFT + Heterogenous Prototype	Cosine similarity
	Cai *et al.* [196]	Coupled least squares regression method with a local consistency constraint	
	Tsai *et al.* [170]	Domain adaptation based proposed DiCA	Subject-specific SVM
	Lin *et al.* [97]	Affine transformations	CNNs over Mahalanobis and Cosine scores
	Fu *et al.*[54]	Designed a dual variational generatorto learn the joint distribution of paired images	
Composite	Chugh *et al.* [33]	Histogram of image moments and HoG	Chi-squared distance
	Han *et al.* [64]	MLBP of ASM features	Similarity on normalized histogram intersection
	Mittal *et al.* [117]	Deep Boltzmann Machines	Neural Networks
	Mittal *et al.* [116]	HoG + DAISY	Chi-squared distance + Attribute feedback
Both	Klum *et al.* [87]	SketchID- automated system based on holistic [83] and component [64] based algorithms	
	Ouyang *et al.*[130]	Learned a mapping to reverse the forgetting process of the eyewitness	
	Zheng *et al.*[213]	Knowledge graph model trained using meta-continual learning	



(a) Viewed sketches (b) Forensic sketches

Figure 2-1: Illustrating the variations in the information content of digital images and different types of sketches.

sic hand-drawn sketch and corresponding photo pairs. Along with domain differences, variations caused by eyewitness description makes this problem further challenging.

Traditionally, a sketch image is matched with digital mugshot images for identifying the suspect. The literature is spread across hand-drawn and composite sketch to digital photo matching [124], with algorithms being evaluated [86, 97] on viewed sketches [87, 185, 209]. Viewed sketches are drawn while looking at the digital photos. Such sketches do not reflect real scenario and fail to capture the challenging nature of the problem. Choi *et al.* [32] have established the limitations of viewed sketches and emphasized the need for new databases and algorithms imitating real scenarios.

Table 2.1 summarizes the literature of facial sketch recognition which shows that both hand-crafted and learned representation have been explored. Sketch recognition has traditionally been viewed as a domain adaptation task due to the cross-domain data. Such techniques can be applied for viewed sketch recognition, where the variations across different types of images is primarily governed by the changes in the domain. However, in case of forensic sketch matching for face images, there are several factors apart from the difference in domain which make the problem further challenging, such as memory gap [130] and the bias observed due to the eye-witness [117]. In this work, we propose a novel transform learning based formulation, *DeepTransformer*, which learns meaningful coupled representations for sketch and digital images. Further, two important

and challenging application scenarios are used for performance evaluation: (i) age separated digital to sketch matching (both composite and hand-drawn) and (ii) sketch to sketch matching. The effectiveness of the proposed formulation is evaluated on hand-drawn and forensic sketch databases, including a novel sketch database. The key contributions are:

- This is the first work incorporating the concept of Deep Learning in Transform Learning framework. Specifically, novel deep coupled transform learning formulations, *Semi-Coupled* and *Symmetrically-Coupled Deep Transform Learning*, have been presented which imbibe qualities of deep learning with domain adaption.

- This is the first work which presents sketch to sketch matching as an important, yet unattended application for law enforcement. As shown in Figure 2-1, composite and hand-drawn sketches have significant difference in their information content. Such matching can be useful for crime linking, where different methods may have been used to generate the sketches.

- IIIT-D CSA dataset [1] contains age-separated images of an individual against a sketch image, for 150 subjects. The dataset also contains 3529 digital images.

2.2 Preliminaries

Dictionary Learning has been used in literature to learn filters and feature representations [95, 127]. For a given input $\mathbf{X}$, a dictionary $\mathbf{D}$ is learned along with the coefficients $\mathbf{Z}$:

$$\min_{\mathbf{D},\mathbf{Z}} \|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_F^2, \ such\ that\ \ \|\mathbf{Z}\|_0 \leq \tau \tag{2.1}$$

where, the l_o-norm imposes a constraint of sparsity on the learned coefficients. It can be observed that dictionary learning is a synthesis formulation; i.e., the learned coefficients and dictionary are able to $synthesize$ the given input $\mathbf{X}$. Ravishankar and Bresler [146] proposed it's analysis equivalent, termed as transform learning. It analyzes the data by learning a transform or basis to

[1] Dataset is available at www.iab-rubric.org/resources/csa.html

produce coefficients. Mathematically, for input data $\mathbf{X}$, it can be expressed as:

$$\min_{\mathbf{T},\mathbf{Z}} \|\mathbf{TX} - \mathbf{Z}\|_F^2 \,,\ such\ that\ \|\mathbf{Z}\|_0 \leq \tau \tag{2.2}$$

where, $\mathbf{T}$ and $\mathbf{Z}$ are the transform and coefficients, respectively. Relating transform learning to the dictionary learning formulation in Equation 2.1, it can be seen that dictionary learning is an inverse problem while transform learning is a forward problem. In order to avoid the degenerate solutions of Equation 2.2, the following formulation is proposed [146]:

$$\min_{\mathbf{T},\mathbf{Z}} \|\mathbf{TX} - \mathbf{Z}\|_F^2 + \lambda \left(\epsilon \|\mathbf{T}\|_F^2 - log \det \mathbf{T}\right) s.t.\ \|\mathbf{Z}\|_0 \leq \tau \tag{2.3}$$

The factor '$log \det \mathbf{T}$' refers to the log-determinant regularizer [91], which imposes a full rank on the learned transform to prevent degenerate solutions. The additional penalty term $\|\mathbf{T}\|_F^2$ is to balance scale. In literature, an alternating minimization approach has been presented [147, 148] to solve the above transform learning problem, i.e.:

$$\mathbf{Z} \leftarrow \min_{\mathbf{Z}} \|\mathbf{TX} - \mathbf{Z}\|_F^2 \,,\ such\ that\ \ \|\mathbf{Z}\|_0 \leq \tau \tag{2.4}$$

$$\mathbf{T} \leftarrow \min_{\mathbf{T}} \|\mathbf{TX} - \mathbf{Z}\|_F^2 + \lambda \left(\epsilon \|\mathbf{T}\|_F^2 - log \det \mathbf{T}\right) \tag{2.5}$$

The coefficients in Equation 2.4 are updated using Orthogonal Matching Pursuit (OMP) [134], and transform matrix $\mathbf{T}$ is updated using a closed form solution presented in [150]. The proof for convergence of the update algorithm can be found in [148]. There is a computational advantage of transform learning over dictionary learning. The latter is a synthesis formulation, and during the test stage, for a given x_{test} it needs to solve a problem of the form:

$$\min_{z_{test}} \|x_{test} - \mathbf{D}z_{test}\|_F^2 \,,\ such\ that\ \ \|z_{test}\|_0 \leq \tau \tag{2.6}$$

This is an iterative optimization problem, and thus time consuming, whereas, transform learning is an analysis framework, and at testing time, only the given equation is solved:

$$\min_{z_{test}} \|\mathbf{T}x_{test} - z_{test}\|_F^2, \; such\ that \; \|z_{test}\|_0 \leq \tau \tag{2.7}$$

This can be solved using one step of hard thresholding [19], making test feature generation very fast and real time.

2.3 DeepTransformer: Proposed Coupled Deep Transform Learning

Transform Learning has been used for several applications such as blind compressive sensing, online learning, along with image and video de-noising [139, 149, 150]. This research addresses the challenging task of sketch recognition by proposing two novel formulations: semi-coupled, and symmetrically-coupled transform learning. This is the first work which incorporates a mapping function in the transform learning framework in order to reduce between-domain variations. Further, both the models have been extended to propose Semi-Coupled DeepTransformer and Symmetrically-Coupled DeepTransformer.

2.3.1 Semi-Coupled Deep Transform Learning

As a result of varying information content of images belonging to different domains, there is a need to reduce the domain gap while performing recognition. This is often achieved by mapping the information content of one domain's data onto the other. In real world scenarios of photo to sketch matching, generally a probe sketch image is matched with a gallery of mugshot digital images. This presents the requirement of transforming data from one domain (sketch) onto the other (digital image). For such instances, where the data from only one domain is required to be mapped to the other, Semi-Coupled Transform Learning is proposed. Let $\mathbf{X_1}$ be the data of first domain and $\mathbf{X_2}$ be the data of second domain. The proposed model learns two transform matrices, $\mathbf{T_1}$ and $\mathbf{T_2}$ (one for each domain) and their corresponding features $\mathbf{Z_1}$ and $\mathbf{Z_2}$, such that

the features from the first domain can be linearly mapped (M) onto the other. Mathematically this is expressed as:

$$\begin{aligned}\min_{\mathbf{T_1},\mathbf{T_2},\mathbf{Z_1},\mathbf{Z_2},\mathbf{M}} &\|\mathbf{T_1X_1}-\mathbf{Z_1}\|_F^2+\ \|\mathbf{T_2X_2}-\mathbf{Z_2}\|_F^2\\ +\lambda\big(\epsilon\,\|\mathbf{T_1}\|_F^2+\ \epsilon\,\|\mathbf{T_2}\|_F^2-\ &log\det\mathbf{T_1}-\ log\det\mathbf{T_2}\big)\\ &+\mu\,\|\mathbf{Z_2}-\mathbf{MZ_1}\|_F^2\end{aligned} \tag{2.8}$$

Equation 2.8 is solved using alternating minimization approach. Specifically, this equation can be decomposed into five sub-problems, one for each variable, and then each is solved individually, as explained below.

Sub-Problem 1:

$$\min_{\mathbf{T_1}}\|\mathbf{T_1X_1}-\mathbf{Z_1}\|_F^2+\ \lambda(\epsilon\,\|\mathbf{T_1}\|_F^2-\ log\det\mathbf{T_1}) \tag{2.9}$$

Sub-Problem 2:

$$\min_{\mathbf{T_2}}\|\mathbf{T_2X_2}-\mathbf{Z_2}\|_F^2+\ \lambda(\epsilon\,\|\mathbf{T_2}\|_F^2-\ log\det\mathbf{T_2}) \tag{2.10}$$

The solution for Equations 2.9, 2.10 is similar to the one for Equation 2.5.

Sub-Problem 3:

$$\begin{aligned}&\min_{\mathbf{Z_1}}\|\mathbf{T_1X_1}-\mathbf{Z_1}\|_F^2+\ \mu\,\|\mathbf{Z_2}-\mathbf{MZ_1}\|_F^2\\ &\equiv\ \min_{\mathbf{Z_1}}\left\|\begin{pmatrix}\mathbf{T_1X1}\\ \sqrt{\mu}\mathbf{Z_2}\end{pmatrix}-\begin{pmatrix}\mathbf{I}\\ \sqrt{\mu}\mathbf{M}\end{pmatrix}\mathbf{Z_1}\right\|_F^2\end{aligned} \tag{2.11}$$

Sub-Problem 4:

$$\begin{aligned}&\min_{\mathbf{Z_2}}\|\mathbf{T_2X_2}-\mathbf{Z_2}\|_F^2+\ \mu\,\|\mathbf{Z_2}-\mathbf{MZ_1}\|_F^2\\ &\equiv\ \min_{\mathbf{Z_2}}\left\|\begin{pmatrix}\mathbf{T_2X2}\\ \sqrt{\mu}\mathbf{MZ_1}\end{pmatrix}-\begin{pmatrix}\mathbf{I}\\ \sqrt{\mu}\,\mathbf{I}\end{pmatrix}\mathbf{Z_2}\right\|_F^2\end{aligned} \tag{2.12}$$

The above two equations are least square problems with a closed form solution, and thus can be minimized for feature representations $\mathbf{Z_1}$ and $\mathbf{Z_2}$.

Sub-Problem 5:

$$\min_{\mathbf{M}}\|\mathbf{Z_2}-\mathbf{MZ_1}\|_F^2 \tag{2.13}$$

Finally, a mapping M is learned between the representations $\mathbf{Z_1}$ and $\mathbf{Z_2}$ by solving the above least square equation.

Inspired by the success of deep learning [65, 90, 160] to model high level abstractions and learn large variations in data, this research introduces *deep* transform learning. For a k-layered architecture, Semi-Coupled DeepTransformer can be expressed as:

$$\begin{aligned}\min_{\theta}\Big[\sum_{j=1}^{k}\Big(\|\mathbf{T_1^j I_1^j}-\mathbf{Z_1^j}\|_{\mathbf{F}}^{\mathbf{2}}+\|\mathbf{T_2^j I_2^j}-\mathbf{Z_2^j}\|_{\mathbf{F}}^{\mathbf{2}}+\\ +\lambda\big(\epsilon\|\mathbf{T_1^j}\|_F^2+\epsilon\|\mathbf{T_2^j}\|_F^2-\ log\det\mathbf{T_1^j}-\ log\det\mathbf{T_2^j}\big)\Big)+\\ \|\mathbf{Z_2^k}-\mathbf{M}\mathbf{Z_1^k}\|_{\mathbf{F}}^{\mathbf{2}}\Big]\end{aligned} \tag{2.14}$$

where, $\theta=\{\forall_{\mathbf{j=1}}^{\mathbf{k}}(\mathbf{T_1^j},\mathbf{T_2^j},\mathbf{Z_1^j},\mathbf{Z_2^j}),\mathbf{M}\}$. ($\mathbf{T_1^j}$, $\mathbf{I_1^j}$, and $\mathbf{Z_1^j}$) and ($\mathbf{T_2^j}$, $\mathbf{I_2^j}$, and $\mathbf{Z_2^j}$) refer to the transform matrix, input, and learned representations of the j^{th} layer for the two domains respectively. $\mathbf{M}$ refers to the learned linear mapping between the final representations of the k^{th} layer ($\mathbf{Z_1^k}$, $\mathbf{Z_2^k}$). The input to the model, $\mathbf{I_1^1}$ and $\mathbf{I_2^1}$ are $\mathbf{X_1}$ and $\mathbf{X_2}$, i.e. training data of the first and second domains, respectively. For subsequent layers, $\mathbf{I_1^j}$ and $\mathbf{I_2^j}$ correspond to the feature representations learned at the previous layers, i.e. $\mathbf{Z_1^{j-1}}$ and $\mathbf{Z_2^{j-1}}$ respectively. As we go deeper and increase the value of k, Equation 2.14 can be solved similar to Equation 2.8. The problem can be divided into $(4k)$+1 sub-problems via alternating minimization approach: separate sub-problems for solving the transform matrices $(2k)$, and the learned representations $(2k)$, and one for the final mapping $\mathbf{M}$. However, solving $(4k)$+1 sub-problems can be computationally expensive as the number of parameters is large. As a cost effective alternative, the proposed model can be learned with greedy layer-wise optimization. Here we explain the layer-wise optimization for a 2-layered semi-coupled deep transform learning model (similar greedy layer-wise optimization can be followed for $k>2$).

Layer One: Learn the first layer transform matrices ($\mathbf{T_1^1}$, $\mathbf{T_2^1}$) for both domains, along with the representations of the input data ($\mathbf{Z_1^1}$, $\mathbf{Z_2^1}$):

$$\min_{\mathbf{T_1^1},\mathbf{Z_1^1}}\left\|\mathbf{T_1^1X_1}-\mathbf{Z_1^1}\right\|_F^2+\ \lambda\big(\epsilon\left\|\mathbf{T_1^1}\right\|_F^2-log\det\mathbf{T_1^1}\big) \tag{2.15a}$$

$$\min_{\mathbf{T_2^1},\mathbf{Z_2^1}}\left\|\mathbf{T_2^1X_2}-\mathbf{Z_2^1}\right\|_F^2+\ \lambda\big(\epsilon\left\|\mathbf{T_2^1}\right\|_F^2-log\det\mathbf{T_2^1}\big) \tag{2.15b}$$

Layer Two: Using the representations learned in the first layer as input, semi-coupled transform

learning is applied at the second layer to obtain the transform matrices for the second layer, for both domains ($\mathbf{T_1^2}, \mathbf{T_2^2}$):

$$\begin{aligned}\min_{\mathbf{T_1^2},\mathbf{T_2^2},\mathbf{Z_1^2},\mathbf{Z_2^2},\mathbf{M}} & \left\|\mathbf{T_1^2 Z_1^1} - \mathbf{Z_1^2}\right\|_F^2 + \left\|\mathbf{T_2^2 Z_2^1} - \mathbf{Z_2^2}\right\|_F^2 \\ + \lambda\big(\epsilon \left\|\mathbf{T_1^2}\right\|_F^2 + \epsilon \left\|\mathbf{T_2^2}\right\|_F^2 - & \log\det \mathbf{T_1^2} - \log\det \mathbf{T_2^2}\big) \\ & + \mu \left\|\mathbf{Z_2^2} - \mathbf{M Z_1^2}\right\|_F^2\end{aligned} \tag{2.16}$$

2.3.2 Symmetrically-Coupled Deep Transform Learning

In real world scenarios, a given sketch image may be matched with a dataset of different type of sketches for crime-linking. In such cases, learning a single mapping function using semi-coupled transform learning may not be useful. For such cases, symmetrically-coupled transform learning is proposed, where two linear maps are learned; one from the first domain to the second one, and the other from the second domain to the first one. This leads to the following formulation:

$$\begin{aligned}\min_{\mathbf{T_1},\mathbf{T_2},\mathbf{Z_1},\mathbf{Z_2},\mathbf{M_1},\mathbf{M_2}} & \left\|\mathbf{T_1 X_1} - \mathbf{Z_1}\right\|_F^2 + \left\|\mathbf{T_2 X_2} - \mathbf{Z_2}\right\|_F^2 \\ + \lambda\big(\epsilon \left\|\mathbf{T_1}\right\|_F^2 + \epsilon \left\|\mathbf{T_2}\right\|_F^2 - & \log\det \mathbf{T_1} - \log\det \mathbf{T_2}\big) \\ & + \mu\big(\left\|\mathbf{Z_2} - \mathbf{M_1 Z_1}\right\|_F^2 + \left\|\mathbf{Z_1} - \mathbf{M_2 Z_2}\right\|_F^2\big)\end{aligned} \tag{2.17}$$

where, $\mathbf{M_2}$ and $\mathbf{M_1}$ correspond to the mapping matrices to transform feature representations of domain two into those of domain one, and vice versa, respectively. As before, with alternating minimization, Equation 2.17 can be optimized with the help of the following sub-problems:

Sub-Problem 1:

$$\min_{\mathbf{T_1}} \left\|\mathbf{T_1 X_1} - \mathbf{Z_1}\right\|_F^2 + \lambda(\epsilon \left\|\mathbf{T_1}\right\|_F^2 - \log\det \mathbf{T_1}) \tag{2.18}$$

Sub-Problem 2:

$$\min_{\mathbf{T_2}} \left\|\mathbf{T_2 X_2} - \mathbf{Z_2}\right\|_F^2 + \lambda(\epsilon \left\|\mathbf{T_2}\right\|_F^2 - \log\det \mathbf{T_2}) \tag{2.19}$$

Updates for the transform matrices ($\mathbf{T_1}, \mathbf{T_2}$) remain the same as shown in Equations 2.9 and 2.10.

Sub-Problem 3:

$$\begin{aligned}\min_{\mathbf{Z_1}} \left\|\mathbf{T_1 X_1} - \mathbf{Z_1}\right\|_F^2 + \mu\big(\left\|\mathbf{Z_2} - \mathbf{M_1 Z_1}\right\|_F^2 \\ + \left\|\mathbf{Z_1} - \mathbf{M_2 Z_2}\right\|_F^2\big)\end{aligned} \tag{2.20}$$

Sub-Problem 4:

$$\min_{\mathbf{Z_2}} \|\mathbf{T_2 X_2} - \mathbf{Z_2}\|_F^2 + \mu\big(\|\mathbf{Z_2} - \mathbf{M_1 Z_1}\|_F^2 + \|\mathbf{Z_1} - \mathbf{M_2 Z_2}\|_F^2 \big) \tag{2.21}$$

The above two equations for learning the representations $(\mathbf{Z_1}, \mathbf{Z_2})$ of the two domains are least square minimizations, and thus have closed form solutions.

Sub-Problem 5:

$$\min_{\mathbf{M_1}} \|\mathbf{Z_2} - \mathbf{M_1 Z_1}\|_F^2 \tag{2.22}$$

Sub-Problem 6:

$$\min_{\mathbf{M_2}} \|\mathbf{Z_1} - \mathbf{M_2 Z_2}\|_F^2 \tag{2.23}$$

Similar to Equation 2.13, mappings $(\mathbf{M_1}, \mathbf{M_2})$ can be learned by solving the above using least square minimization. As discussed, the DeepTransfomer is solved using alternating minimization approach. Each of the subproblems of Equations 2.8 and 2.17 are solved with guaranteed convergence [146]. Specifically, learning Z has analytical solution and transform updates are done by conjugate gradients which can only decrease. Overall, the model has monotonically decreasing cost function and therefore, will converge.

We extend Equation 2.17 and propose symmetrically-coupled *deep* transform learning where, $\theta = \{\forall_{\mathbf{j=1}}^{\mathbf{k}}(\mathbf{T_1^j}, \mathbf{T_2^j}, \mathbf{Z_1^j}, \mathbf{Z_2^j}), \mathbf{M_1}, \mathbf{M_2}\}$, and $\mathbf{M_2}$ and $\mathbf{M_1}$ correspond to the mapping matrices to transform feature representations of domain two into those of domain one, and vice versa, respectively. It is mathematically expressed as:

$$\begin{aligned}\min_{\theta} \Big[\sum_{j=1}^{k} \Big(\|\mathbf{T_1^j I_1^j} - \mathbf{Z_1^j}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{T_2^j I_2^j} - \mathbf{Z_2^j}\|_{\mathbf{F}}^{\mathbf{2}} + \\ + \lambda\big(\epsilon\|\mathbf{T_1^j}\|_F^2 + \epsilon\|\mathbf{T_2^j}\|_F^2 - \ log\det \mathbf{T_1^j} - \ log\det \mathbf{T_2^j}\big)\Big) + \\ \|\mathbf{Z_2^k} - \mathbf{M_1 Z_1^k}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{Z_1^k} - \mathbf{M_2 Z_2^k}\|_{\mathbf{F}}^{\mathbf{2}} \Big]\end{aligned} \tag{2.24}$$

This formulation can be solved using alternating minimization approach with $(4k$+2) sub-problems where the last two sub-problems are related to learning mappings $\mathbf{M_1}$ and $\mathbf{M_2}$. However, like Semi-Coupled DeepTransformer, we optimize symmetrically coupled deep transform algorithm in a greedy layer wise manner. The optimization for a 2-layer Symmetrically-Coupled DeepTrans-

former is as follows:

Layer One:

$$\min_{\mathbf{T_1^1}, \mathbf{Z_1^1}} \left\|\mathbf{T_1^1 X_1} - \mathbf{Z_1^1}\right\|_F^2 + \lambda\left(\epsilon \left\|\mathbf{T_1^1}\right\|_F^2 - log\det \mathbf{T_1^1}\right) \tag{2.25a}$$

$$\min_{\mathbf{T_2^1}, \mathbf{Z_2^1}} \left\|\mathbf{T_2^1 X_2} - \mathbf{Z_2^1}\right\|_F^2 + \lambda\left(\epsilon \left\|\mathbf{T_2^1}\right\|_F^2 - log\det \mathbf{T_2^1}\right) \tag{2.25b}$$

Layer Two:

$$\begin{aligned}\min_{\mathbf{T_1^2}, \mathbf{T_2^2}, \mathbf{Z_1^2}, \mathbf{Z_2^2}, \mathbf{M}} & \left\|\mathbf{T_1^2 Z_1^1} - \mathbf{Z_1^2}\right\|_F^2 + \left\|\mathbf{T_2^2 Z_2^1} - \mathbf{Z_2^2}\right\|_F^2 \\ + \lambda\left(\epsilon \left\|\mathbf{T_1^2}\right\|_F^2 + \epsilon \left\|\mathbf{T_2^2}\right\|_F^2 - log\det \mathbf{T_1^2} - log\det \mathbf{T_2^2}\right) & \\ + \mu(\left\|\mathbf{Z_2^2} - \mathbf{M_1 Z_1^2}\right\|_F^2 + \left\|\mathbf{Z_1^2} - \mathbf{M_2 Z_2^2}\right\|_F^2) &\end{aligned} \tag{2.26}$$

The first layer learns the low level representation of each domain independently, while the second layer learns the high level representations and mapping between the representations of the two domains/modalities. The proposed model thus encodes domain specific features, followed by features incorporating the between-domain variations.

2.3.3 DeepTransformer for Sketch Recognition

The proposed two layer DeepTransformer is used for performing face sketch matching. For semi-coupled DeepTransformer, the following steps are performed:

Training: Given a set of sketch and digital training pairs, $\mathbf{X_s}, \mathbf{X_d}$, transform matrices $(\mathbf{T_s^1}, \mathbf{T_d^1}, \mathbf{T_s^2}, \mathbf{T_d^2})$ and coefficient vectors $(\mathbf{Z_s^1}, \mathbf{Z_d^1}, \mathbf{Z_s^2}, \mathbf{Z_d^2})$ are learned using Equation 2.14, along with a mapping, $\mathbf{M}$, between $\mathbf{Z_s^2}, \mathbf{Z_d^2}$. A two hidden layer neural network classifier is trained to make identification decisions.

Testing: For a given probe sketch image, x_{sTest}, the first and second layer feature representations are extracted using the learned transform spaces:

$$z_{sTest}^1 = \mathbf{T_s^1} x_{sTest};\ z_{sTest}^2 = \mathbf{T_s^2} z_{sTest}^1 \tag{2.27}$$

The mapping $\mathbf{M}$ is used to transform the feature vector onto the digital image space, i.e, $z_{dTest}^2 = \mathbf{M_1} z_{sTest}^2$. The feature representation of the sketch (probe) in the digital image feature space,

z_{dTest} is now used for performing recognition using the trained neural network. For sketch to sketch matching (i.e. cases where mappings to and from different modalities are required), similar steps can be followed for utilizing symmetrically-coupled deep transform learning.

2.4 Databases and Experimental Protocol

Face sketch databases [87, 185] generally comprise of viewed sketches, either hand-drawn or composite. Viewed sketches are created by looking at a digital image and sketching it simultaneously. This fails to capture the uncertainty in the recall process that humans encounter or the variations in characteristics, like a hairstyle modification, that are generally present between a sketch and digital image acquired at different times. In this research, we utilize a novel IIIT- D *Composite Sketch with Age variations (CSA)* dataset [34], which is the **first** publicly available dataset containing multiple age-separated digital images for a given sketch image. Inspired by Bhatt *et al.* [18], the human forgetting process is incorporated by creating semi-forensic composite sketches. The user is shown the digital image of a subject for a few minutes, and is asked to create the composite sketch after a period of 30 minutes based on his/her memory. The database consists of 3529 sketches and digital face images pertaining to 150 individuals. Out of the 150 subjects, 52 are selected from the FG-NET Aging Database [92], 82 are selected from IIIT-D Aging Database [198], and the remaining subjects are collected from the Internet. The composite sketch images are created using FACES [2], a popular software to generate photo-like composite sketches.

In IIIT-D CSA dataset, the digital images span over an age range of 1 to 65 years. For each subject, an image is chosen from the middle of his/her age range and a corresponding sketch image is generated. Following this, each subject's digital images are divided into three categories:

(i) Younger age group: This category models the scenario when the digital images (gallery) are younger than the probe sketch image. This set contains 1713 digital images.

(ii) Same age group: This category represents the scenario when the age of an individual is similar in both digital image (gallery) and sketch image (probe). A total of 150 digital images exist in this set.

(iii) Older age group: This category imitates the scenario when the digital images (gallery) of the individuals are at an age older than the sketch. It consists of 1516 digital images.

Table 2.2: Details of experimental protocols. For P1-P5, testing is performed on the IIIT-D CSA dataset. For P6-P7, unseen training and testing partitions are used from the CUFS and e-PRIP datasets (both contain sketches pertaining to AR dataset).

Prtcl.	Gallery Type	Probe (Sketch)	Databases for Feature Learning	Test Database	Training Pairs	Gallery Size	Probe Size
Matching sketch to age-separated digital images: Semi-Coupled DeepTransformer							
P1	Younger age	Composite	CUFS, CUFSF, e-PRIP, PRIP-VSGC, IIIT-D Viewed and Semi-viewed	CSA	2129	875	90
P2	Same age	Composite		CSA	2129	90	90
P3	Older age	Composite		CSA	2129	1044	90
P4	Large-scale dataset	Composite		CSA	2129	7165	90
P5		Forensic		IIIT-D Forensic	2129	7265	190
Sketch to sketch matching: Symmetrically-Coupled DeepTransformer							
P6	Composite	Hand-drawn	CUFS, e-PRIP	CUFS, e-PRIP	50	73	73
P7	Hand-drawn	Composite	CUFS, e-PRIP	CUFS, e-PRIP	50	73	73

Overall, IIIT-D CSA consists of 150 composite sketch images, one for each subject, and 3379 digital images belonging to different age categories. Apart from IIIT-D CSA, we have also used viewed hand-drawn sketch and digital image pairs from CUHK Face Sketch Dataset (**CUFS**) [185] (311 pairs of students and AR dataset [109]), CUHK Face Sketch FERET Dataset (**CUFSF**) [209] (1194 pairs), and **IIIT-D Sketch dataset** [18]. IIIT-D dataset contains viewed (238 pairs), semi-viewed (140 pairs), and forensic sketches (190 pairs). Composite sketches from PRIP Viewed Software-Generated Composite database (**PRIP-VSGC**) [87] and extended-PRIP Database (**e-PRIP**) [117] (Indian user set) are also used.

Experimental Protocol: To evaluate the efficacy of the proposed formulations two challenging problems are considered: sketch matching against age-separated digital images (semi-coupled DeepTransformer) and sketch to sketch matching (symmetrically coupled DeepTransformer). Since this is the first research that focuses on sketch to sketch matching, as well as sketch to age-separated digital matching, we have created seven different experimental protocols to understand the performance with individual cases. These protocols are classified according to the two case studies and the details are summarized in Table 2.2.

1. Matching Sketch to Age-Separated Digital Images: CSA test set and IIITD Forensic hand-drawn database have been used to evaluate the performance of the proposed model. Inspired from real life scenarios, the test set is divided into a gallery and probe set. The gallery contains the digital images while the probe contains the sketch image. The first three protocols evaluate the effect of age difference on the recognition performance, and the next two protocols (P4 and P5) analyze the difference in performance on matching forensic and composite sketches with large scale digital image gallery. Since sketch to digital image matching experiment involves one way mapping, the results are demonstrated with Semi-Coupled DeepTransformer.

2. Sketch to Sketch Matching: In real world crime scene linking application, one might want to match a hand-drawn sketch against a database of composite sketches, or the other way around. Therefore, for this experiment, the proposed Symmetrically-Coupled DeepTransformer is used. CUFS dataset contains hand-drawn sketch images for the AR dataset (123 subjects), while e-PRIP contains composite sketches generated by a sketch artist for the same. The following two experiments with protocols P6 and P7 are performed: **(i)** composite to hand-drawn sketch, and **(ii)** hand-drawn to composite sketch matching.

2.5 Results and Observations

Effectiveness of DeepTransformer is evaluated with multiple input features, namely Dense Scale Invariant Feature Transform (DSIFT) [21], Dictionary Learning (DL) [95], Class Sparsity based Supervised Encoder (L-CSSE) [106], Light CNN [192], and VGG-Face [133]. To analyze the effect of depth in this formulation, the results are computed with single layer (low level features) and with two layers (high level features) of DeepTransformer. Two kinds of comparative experiments are performed. The first one compares the performance of one layer and two layers deep transform learning algorithms with two classifiers, i.e., Euclidean distance and neural network. The second comparison is performed with existing algorithms like Semi-Coupled Dictionary Learning algorithm (SCDL) [183] and Multi-Modal Sharable and Specific feature learning algorithm (MMSS) [175]. Both the techniques have been used in literature for performing cross-domain recognition, wherein the former is a coupled dictionary learning based approach (synthesis technique), and the latter incorporated transform learning with convolutional neural networks for addressing

Table 2.3: Rank-10 accuracies (%) for protocols P1 to P3 using proposed Semi-Coupled DeepTransformer.

Features	Euclidean Distance	NNET	DeepTransformer	
			1-Layer	2-Layer
Gallery with Younger Age Digital Images (P1)				
DSIFT	8.9	15.6	26.7	27.8
DL	2.2	14.4	17.8	17.8
VGG	1.1	11.1	12.2	12.2
Light CNN	8.9	12.2	30.0	27.8
L-CSSE	14.4	19.7	34.1	**42.6**
Gallery with Same Age Digital Images (P2)				
DSIFT	7.8	26.7	25.6	27.8
DL	1.1	13.3	15.6	17.8
VGG	2.2	12.2	14.4	14.4
Light CNN	11.1	25.6	32.2	34.4
L-CSSE	16.3	30.2	37.7	**44.2**
Gallery with Older Age Digital Images (P3)				
DSIFT	5.6	21.1	23.3	24.4
DL	2.2	13.3	17.8	18.9
VGG	2.2	11.1	12.2	12.2
Light CNN	7.8	20.0	24.4	28.9
L-CSSE	9.9	20.0	28.9	**36.0**

cross-domain recognition. Comparison has also been drawn with state-of-the-art sketch recognition algorithms, namely MCWLD [18] and GSMFL [97], and a commercial-off-the-shelf system (COTS), Verilook [4]. In all the experiments, for training the networks, data augmentation is performed on the gallery images to increase per-class samples by varying the illumination and flipping the images along the y-axis. The key observations from experimental results are:

Performance with Different Features: Tables 2.4 and 2.3 present the rank-10 identification accuracies of DeepTransformer with different features, for both applications of sketch matching: sketch to sketch matching and sketch to photo matching. Table 2.3 presents the accuracies for sketch to digital image matching, where the proposed Semi-Coupled DeepTransformer has been used. The results show that DeepTransformer enhances the performance of existing feature extraction techniques by at least 10% as compared to Euclidean distance matching, and at most 22% when neural network (NNET) is used for classification. Upon comparing accuracies across features, it is ob-

served that DeepTransformer achieves the best results with L-CSSE features for all protocols. Similar results can be seen from Table 2.4 where Symmetrically-Coupled DeepTransformer has been used for sketch to sketch matching. Experimentally, it can be observed that providing class-specific features to DeepTransformer results in greater improvement. L-CSSE is a supervised deep learning model built over an autoencoder. The model incorporates supervision by adding a $l_{2,1}$ norm regularizer during the feature learning to facilitate class-specific feature learning. The model utilizes both global and local facial regions to compute feature vector and has been shown to achieve improved results for existing face recognition problems. Further, we also observe that L-CSSE encodes the high frequency features in both local and global regions which are pertinent to digital face to sketch matching. Moreover, improved performance is observed for hand-crafted, as well as representation learning based features, thus promoting the use of DeepTransformer for different types of feature extraction techniques and input data.

Comparison with Existing Approaches: Table 2.5 shows that the proposed, DeepTransformer with L-CSSE features outperforms existing algorithms for both the applications of sketch recognition. In case of sketch to digital image matching, with younger age protocol (P1), Semi-Coupled DeepTransformer attains a rank-10 accuracy of **42.6%**, which is at least 15% better than existing algorithms, and around 24% better than COTS. Similar trends are observed for P2 and P3 protocols, where the proposed DeepTransformer outperforms existing techniques and the commercial-off-the-shelf system by a margin of at least 13% and 11% respectively. Additionally, the matching accuracy achieved by the proposed Symmetrically-Coupled DeepTransformer exceeds existing techniques for the task of sketch to sketch matching as well (P6, P7). An improvement of at least 14% and at most 20% is seen with the proposed DeepTransformer (L-CSSE as feature) for the given protocols. This accentuates the use of DeepTransformer for addressing the problem of real world sketch matching.

Effect of Layer-by-Layer Training and Number of Layers: We compare the performance of the proposed DeepTransformer with and without layer-by-layer training (i.e. Equations 2.14 and 2.24 for direct solving for $k = 2$ and layer-by-layer training as per Equations 2.15-2.16 and 2.25-2.26). On a 108-core server with 256GB RAM, for protocols P1 to P3, training Semi-Coupled Deep-

Table 2.4: Rank-10 accuracies (%) for sketch to sketch matching (P6, P7) using Symmetrically-Coupled DeepTransformer.

	Gallery: Composite Probe: Hand-drawn (P6)				**Gallery: Hand-drawn Probe: Composite (P7)**			
Features	**Euclidean Distance**	**NNET**	**DeepTransformer**		**Euclidean Distance**	**NNET**	**DeepTransformer**	
			1-Layer	**2-Layer**			**1-Layer**	**2-Layer**
DSIFT	4.1	16.4	24.7	28.8	2.7	13.7	23.3	30.1
DL	4.1	15.1	17.8	19.2	6.9	16.4	19.2	20.6
VGG	6.9	12.3	24.7	27.4	6.9	15.1	19.2	20.6
Light CNN	8.2	15.1	26.0	30.1	8.2	16.4	20.6	28.8
L-CSSE	10.9	17.8	28.4	**31.5**	10.9	20.9	31.5	**33.6**

Table 2.5: Rank-10 accuracies (%) comparing proposed DeepTransformer with existing algorithms and COTS.

Algorithm	**P1**	**P2**	**P3**	**P6**	**P7**
MCWLD [18]	26.8	30.7	24.4	16.5	19.2
GSMFL [97]	25.2	29.3	23.3	16.5	19.2
SCDL [183]	23.3	25.6	18.9	15.1	13.7
MMSS [175]	22.2	27.8	21.1	13.3	15.1
Verilook (COTS) [4]	17.8	16.6	12.2	10.9	13.7
DeepTransformer (with L-CSSE)	**42.6**	**44.2**	**36.0**	**31.5**	**33.6**

Transformer with layer-by-layer training requires 142 seconds which is 12 seconds faster than without layer-by-layer training. For both the cases, for $k = 2$, the rank-10 accuracies are same which shows that layer-by-layer training is cost effective. We also analyze the effect of number of layers and, as shown in Tables 2.4 and 2.3, 1-8% improvement in rank-10 accuracy is observed for different protocols upon going deeper.

Performance on Large-Scale Dataset: The performance of the proposed DeepTransformer has also been evaluated for a large-scale real world dataset using protocols P4 and P5. Figure 2-2 presents the Cumulative Match Characteristic curves (CMCs) for IIIT-D CSA composite and IIIT-D Forensic hand-drawn sketch database respectively. The proposed Semi-Coupled DeepTransformer achieves a rank-50 accuracy of 33.7%, which is an improvement of at least 5% from other algorithms on IIIT-D CSA dataset. Similar results can be observed on the forensic sketches as

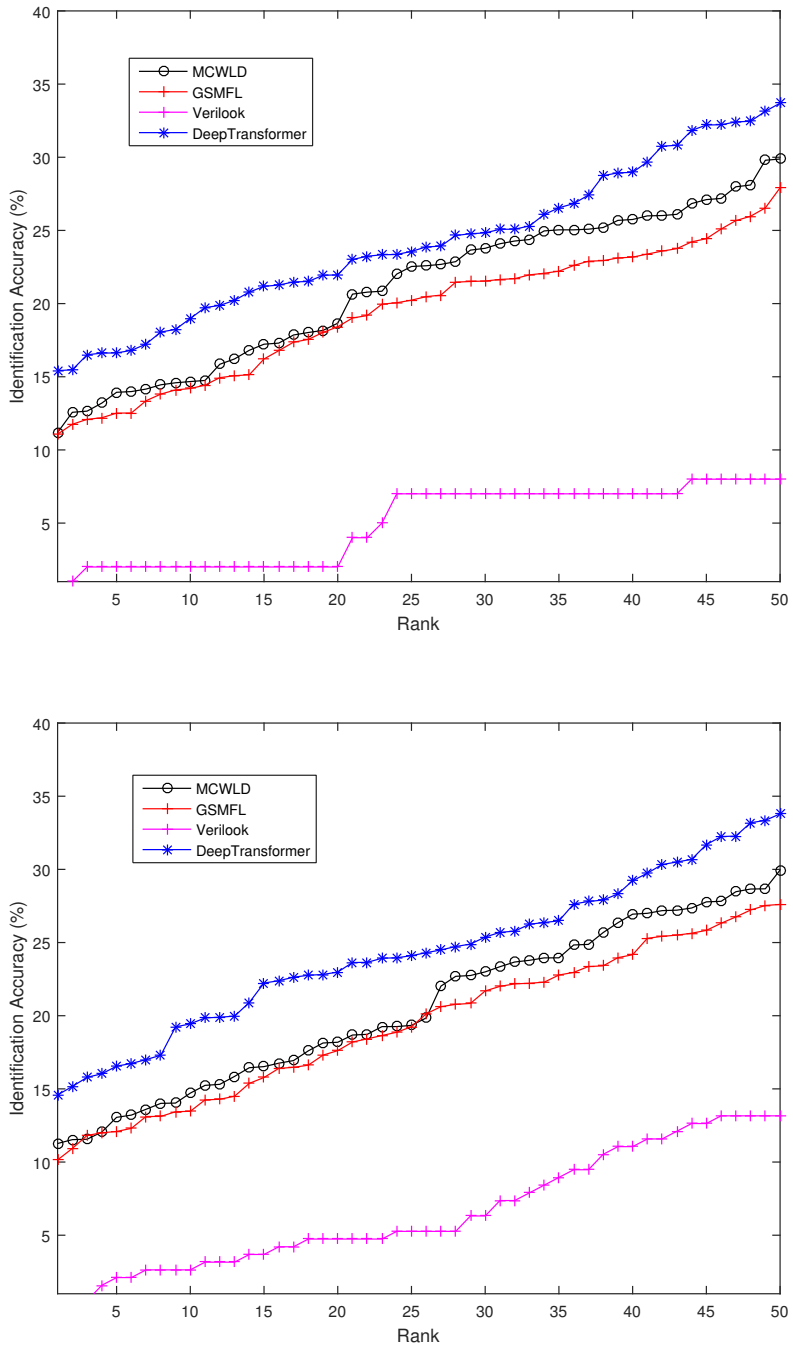


Figure 2-2: CMC curves for P4 and P5 experiments: (a) CSA, and (b) IIIT-D Forensic datasets.

well.

The experimental results showcase the efficacy of the proposed DeepTransformer, in terms of the improvement in identification accuracies with different features, and in comparison with other existing models. The results suggest that the DeepTransformer is robust to the type of feature and

has the ability to learn over varying input spaces. Moreover, efficient training of the symmetrically coupled DeepTransformer with as few as 50 digital-sketch pairs (P6 and P7) motivate the use of the proposed architecture for small sample size problems as well. The evaluation on different real world protocols further strengthens the usage of the proposed model for addressing cross domain matching tasks.

2.6 Summary

This research focuses on the challenging problem of face sketch recognition and proposes a novel transform learning based formulation, called as *DeepTransformer*. Two models: Semi-Coupled and Symmetrically-Coupled DeepTransformer have been presented, both of which aim to reduce the variations between two domains. The highlight of the proposed formulation is that it provides the flexibility of using an existing feature extractor and classifier in the framework. The proposed DeepTransfomer is evaluated with real world scenarios of age-separated digital image to sketch matching and sketch to sketch matching. Results are also shown on the IIIT-D Composite Sketch with Age variations database of 150 subjects. Comparison with existing state-of-the-art algorithms and commercial-off-the-shelf system further instantiates the efficacy of both the semi-coupled and symmetrically coupled variants of the purposed DeepTransformer. This research has been published in the IEEE/CVF International Conference on Computer Vision (ICCV), 2017, and all the figures have been taken from the published research paper [123].

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Discriminative Shared Transform Learning for Sketch to Image Matching

3.1 Introduction

Sketch to digital image matching is a special kind of cross-domain matching task, where given a query instance of domain *A*, samples of the same category are retrieved from a different domain *B* (as illustrated in Figure 3-1). Sketch-to-digital image matching finds applicability in important real-world applications such as surveillance by means of forensic sketch face recognition [86], and social media profile linking or person identification via caricature recognition [111]. With the increasing usage of touchscreens and e-commerce, sketch based image retrieval has become an integral part of image retrieval tasks as well [138, 173, 186, 199]. The given problem suffers from the availability of limited training data, and possesses a unique challenge where the intra-domain similarity often overpowers the intra-class similarity. For example, a sample appears more similar to another class's sample from the same domain, as compared to its own class's sample from a different domain.

In order to address the cross-domain matching task of sketch to image matching, researchers have proposed several algorithms to model the inter-domain variations [129, 177]. Most of the existing techniques have either focused on learning an explicit transformation function from one domain to another, or on learning a shared space for two domains. The first approach involves learning a transformation function between the two domains [123, 182], however, it often requires

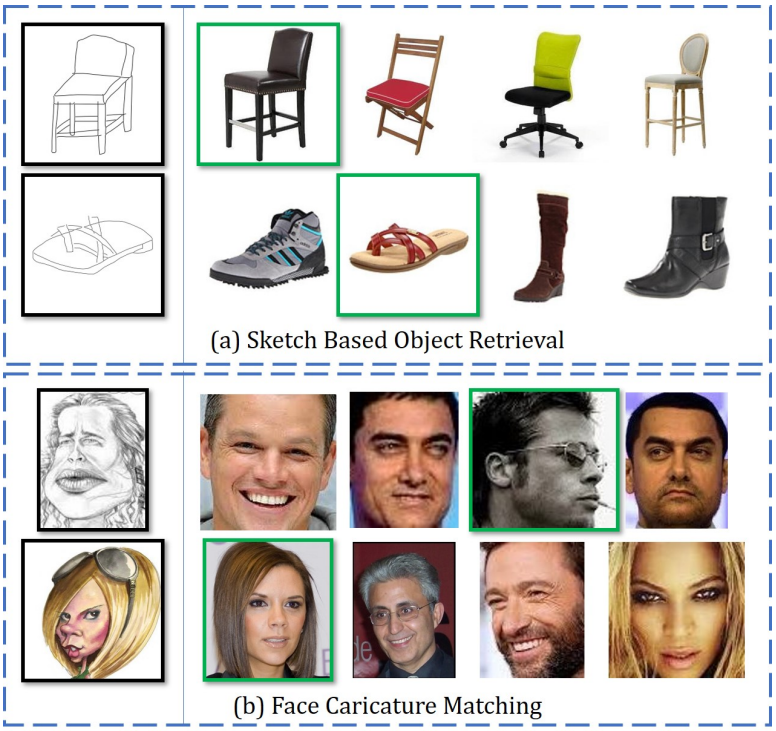


Figure 3-1: Sample applications of cross-domain matching demonstrating variation in the information content. Query from domain A (i.e. sketch domain) is matched against a database of another domain B (i.e. image domain).

large amount of data for modeling the inter-domain variations. For instance, in face recognition applications such as sketch to digital image matching, a learned transformation should be generalizable to new identities seen only during testing. The second approach explicitly models the inter-domain variations by learning a common space for data belonging to two domains [47, 164, 193]. Algorithms belonging to this category learn representations in a shared space, thus minimizing the cross-domain variations. This is often achieved by utilizing *representation learning*, which is able to encode effective features for the input data. Beyond cross-domain variations, the problem of sketch to image matching is further exacerbated with the availability of limited training samples.

This research proposes a novel *Discriminative Shared Transform Learning* algorithm for sketch to digital image matching. The proposed algorithm learns domain invariant and discriminative features under supervised constraints. It is able to learn effective features for cross-domain matching with a small number of parameters, and improve the current state-of-the-art performance. The key highlights of this research are:

- This research presents a novel learning algorithm, *Discriminative Shared Transform Learn-*

ing, for cross-domain matching, specifically sketch to digital image matching. As part of the proposed technique, two models have also been presented: (i) *Contractive model (C-model)*[1], and (ii) *Divergent model (D-model)*. The efficacy of the proposed algorithm is demonstrated on seven datasets pertaining to three sketch to digital image matching case-studies: (i) caricature face recognition, (ii) sketch based object image retrieval, and (iii) sketch face recognition.

- A shared transform is learned which enables learning of domain-invariant features for data belonging to different domains. It also reduces the number of parameters to be learned, thereby making it a suitable choice for scenarios with limited training data.

- The proposed algorithm is feature agnostic, i.e. it can operate on different features as input in order to achieve enhanced performance. Extensive evaluation on the proposed C-model and D-model demonstrates the effectiveness of the proposed algorithm with different input features. Detailed analysis also highlights the importance of each component of the proposed algorithm.

3.2 Related Work

With increase in automated computer vision applications problems related to cross-domain matching, specifically sketch to digital image matching, has been gaining attention of the research community. According to the case studies presented in Figure 3-1, the related work is divided into two subsections: (i) Sketch based Image Retrieval and (ii) Sketch and Caricature Face Recognition.

Sketch based Image Retrieval (SBIR): SBIR refers to the task of retrieving digital images similar to the given sketch query image (generally objects). Since the query and retrieved samples belong to different domains, cross-domain matching is performed. SBIR is relevant in today's context for applications such as online shopping websites, or for searching similar images by sketching the object on our touch screens.

Due to the lack of color and texture information in object sketch images, traditional approaches

[1] A preliminary version of this formulation has been published in IEEE BTAS 2018 [161].

utilized structural information to extract features for matching. Cao *et al.* [26] proposed a real time approach based on structure index, termed as EdgelIndex and a corresponding contour based matching algorithm. Similar to the Bag of Words model, Eitz *et al.* [50] proposed the Bag of Visual Words (BOVW) model for SBIR. BOVW utilizes images as visual words, followed by clustering for generating a codebook, and histogram matching. Hu *et al.* [69] presented a variant of the popular Histogram of Oriented Gradients (HOG) feature extractor, termed as Gradient Field-HOG (GF-HOG), useful for extracting low-level features from sketch and digital images. A combination of GF-HOG and BOVW was presented for improved SBIR performance.

Recently, some deep learning models have also been proposed for sketch based image retrieval. Qi *et al.* [142] presented a Siamese based convolutional neural network for SBIR. Song *et al.* [203] presented a new dataset of shoes and chairs for fine-grained SBIR. The authors proposed a triplet ranking deep model and a novel pre-training strategy. This was followed by a spatially aware attention module [164], which combines fine semantic information along with coarse information for fine-grained SBIR. Recently, Huang *et al.* [71] highlighted and addressed the several challenges faced in SBIR. The authors proposed a deep visual semantic descriptor to encode low level and high level features for both the domains, followed by a clustering based re-ranking approach. Zhang *et al.* [208] present a technique which dynamically discovers landmarks, which aids in learning the discriminative structural representations. Further, Zhang *et al.* [210] proposed a Hybrid CNN model for modeling the appearance and shape information for sketch based image retrieval. Sketch based image object retrieval has also been addressed by utilizing pre-trained deep learning models with domain-specific information [43, 47, 101]. Owing to the large number of parameters, most of the deep learning based models suffer from the major challenge of requiring large amount of labeled training data (either for training or pre-training), and often, labeled data pairs across domains. The data is required to effectively train the model in order to learn efficient representations, often rendering the model unscalable for large number of classes when trained with limited data.

Sketch and Caricature Face Recognition: Ouyang *et al.* [129] documented the existing techniques for sketch based face recognition [52, 136], including the other scenarios of heterogeneous face matching such as infra-red, 3D, and low resolution. For sketch face recognition, past ap-

proaches can be divided into feature based [18, 64, 209], synthesis based [185], and projection based [70, 83]. Feature based approaches focus on extracting domain invariant features for the given data. Synthesis based approaches transform data from one domain to another in order to reduce the inter domain gap, while projection based approaches project data belonging to both the domains onto a common subspace. Since projection based techniques do not transform data from one domain to another, they are often preferred as compared to synthesis based methods. In the past, researchers have also tried to address the memory gap present in sketch face recognition by modeling the human forgetting process [130]. Pereira *et al.* [40] proposed utilizing the high-level features of deep learning models for learning domain specific units. Nagpal *et al.* [191] proposed deep transform learning for sparse feature extraction, followed by a mapping between the representations of the two domains. More recently, Zheng *et al.* [213] present a meta learning and knowledge graph based framework which learns from the relational knowledge of the samples. Fu *et al.* [54] formulates the problem as a dual generator problem and present a novel network framework to generate a pair of heterogenous images from noise. Liu *et al.* [99] develop an iterative re-ranking based for attribute guided synthesized sketches to eliminate bias during sketch synthesis process.

Limited research has been performed to automate the process of face caricature matching. Klare *et al.* [84] utilized 25 qualitative features describing the face features such as the face shape, nose shape, and hair type to encode the image and caricature faces. The authors also created the first publicly available dataset for caricature face recognition. Takayama *et al.* [169] proposed using a similarity vector based on skin color, hair type, and hair quantity for face caricature recognition. Recently, Shi *et al.* [158] proposed a caricature generation technique for creating a caricature from a given digital face image. It is important to note that most of the existing automated caricature face recognition approaches have primarily focused on using annotated qualitative features. This results in the requirement of annotated information for the testing data, which often limits the utility of the recognition algorithm.

Despite the recent advances in sketch to digital image matching, the performance achieved by the state-of-the-art techniques presents a need for further improvement, especially for real world scenarios (eg. 19.44% rank-10 performance on the IIIT-D Forensic sketch dataset [18]). Most of the existing approaches focus on extracting hand-crafted features, useful for particular domains;

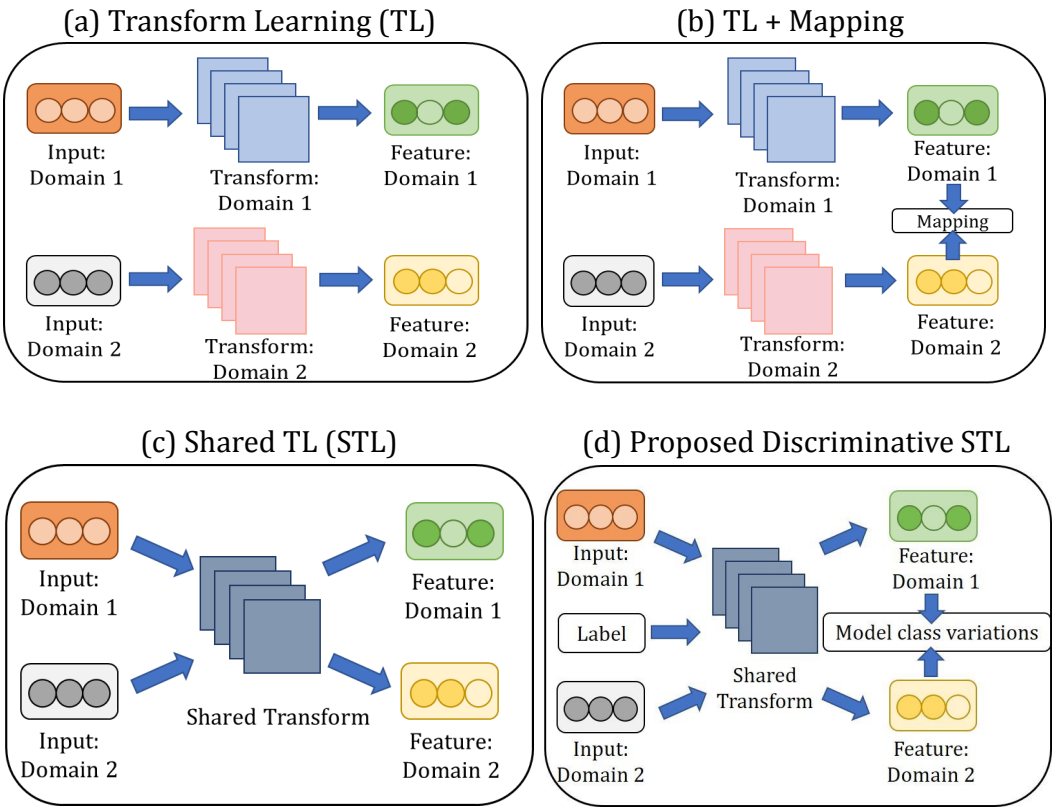


Figure 3-2: Diagrammatic representation of how transform learning can be used for cross-domain matching.

utilize task-specific knowledge (eg. incorporating facial landmarks for face retrieval); or contain large number of trainable parameters requiring ample amount of paired training data from different domains (eg. deep learning based techniques). Recent representation learning techniques transform data from one domain to another or project data onto a common subspace. As mentioned previously, learning a generalized transformation function is often a challenging task, requiring large amount of training data, while projections onto a common subspace may often result in the loss of useful discriminative information.

3.3 Proposed Discriminative Shared Transform Learning

Traditionally, transform learning is not directly applicable to cross-domain matching. In order to use the vanilla transform learning methods for heterogeneous matching, separate transforms are learned for the two domains, followed by matching via a classifier or distance metric (Figure 3-

2(a)). In the literature, transform learning has been extended for cross-domain matching by incorporating a mapping matrix between the learned features (Figure 3-2(b)) [123]. These techniques do not explicitly minimize the inter-domain variations, and require learning multiple transforms, resulting in a large number of learnable parameters, which is often challenging with the limited training data available. This research addresses the above limitations by proposing *Discriminative Shared Transform Learning (DSTL)* to learn domain-invariant features for matching data belonging to different domains, primarily sketch and digital images. The proposed algorithm enables learning of domain invariant features, which encode differentiable information, and is not data intensive, i.e. suitable for limited training data.

3.3.1 DSTL: Components and Formulation

The DSTL algorithm is built using two novel components: (i) shared transform learning, and (ii) discriminative feature learning. Details regarding each are provided in this subsection.

Shared Transform Learning: It involves learning a single transform matrix ($\mathbf{T}$) for data belonging to two domains. A single transform enables the model to learn features common across domains, thereby reducing the inter-domain variations. In sketch-to-digital image matching, this would correspond to projecting the digital and sketch images using the same transform matrix. STL eliminates the need for multiple domain-specific transforms, and facilitates learning of a transform capable of representing data of both the domains. For training samples $\mathbf{X_1}$ and $\mathbf{X_2}$ of two domains, STL is formulated as:

$$\min_{\mathbf{T},\mathbf{Z_1},\mathbf{Z_2}} \|\mathbf{TX_1} - \mathbf{Z_1}\|_F^2 + \|\mathbf{TX_2} - \mathbf{Z_2}\|_F^2 + \lambda\big(\epsilon\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\big) \tag{3.1}$$

where, $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are the learned representations for the corresponding input data, and $\mathbf{T}$ is the shared transform matrix.

Discriminative Feature Learning: Since STL is unsupervised in nature, it does not model the relationship between different classes. To this effect, the second component, i.e. *discriminative feature learning*, is incorporated into the proposed DSTL algorithm. Class information is utilized via discriminative losses, such that the learned features are useful for classification. A distance term, $\mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$ is introduced in the loss function, which aids in learning discriminative fea-

tures by encoding class variations at the time of feature learning.

Given training data pertaining to two different input domains, namely digital image, $\mathbf{X_1}$, and sketch image, $\mathbf{X_2}$, the corresponding representations $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are learned using the proposed DSTL algorithm. Building on Equation 3.1, DSTL is mathematically expressed as follows:

$$\min_{\mathbf{T},\mathbf{Z_1},\mathbf{Z_2}} \|\mathbf{TX_1} - \mathbf{Z_1}\|_F^2 + \|\mathbf{TX_2} - \mathbf{Z_2}\|_F^2 + \lambda\big(\epsilon\,\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\big) + \mu\mathcal{F}(\mathbf{Z_1},\mathbf{Z_2},\mathbf{y}) \tag{3.2}$$

where, λ, μ, and ϵ correspond to the regularization parameters which govern the weight given to different terms.

3.3.2 Variants of the DSTL Algorithm

DSTL enables learning domain invariant features under supervised constraints. The distance term ($\mathcal{F}(.)$) in Equation 3.2 can be modeled for handling the class variations across domains. In sketch to digital image matching, samples of the same class, belonging to different domains, suffer from the challenge of high intra-class variations due to the varying information content in both the input samples. In order to model the class variations across domains, the *Contractive Model (C-model)* of DSTL is proposed to learn features for sketch to digital image matching.

Contractive Model (C-model): The C-model learns a shared transform matrix while reducing the intra-class variations. Here, the input $\mathbf{X_1}$ and $\mathbf{X_2}$ are created such that they contain pair-wise data. That is, each pair contains samples of the same class, belonging to the two domains. Specifically, in case of sketch-to-digital image matching, the i^{th} sample of $\mathbf{X_1}$ and $\mathbf{X_2}$ would belong to class-A, where one would be a sketch image and the other is a digital image. Therefore, C-model does not utilize the class labels (Y) directly into the formulation, however, it uses the class information to create pairs for training. Given same-class (or genuine) pairs only, $\mathcal{F}(\mathbf{Z_1},\mathbf{Z_2},\mathbf{y})$ is modeled as the Euclidean distance between the representations. This corresponds to a similarity-preserving loss which minimizes the distance between the representations of the same class pairs. C-model is thus

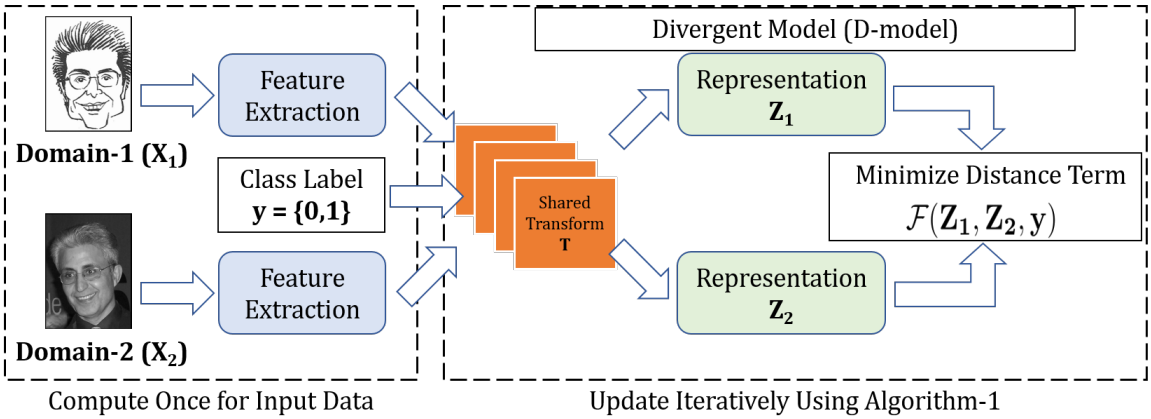


Figure 3-3: The proposed D-model uses cross-domain image pairs along with their labels. D-model learns a shared transform $\mathbf{T}$ while modeling the inter-class and intra-class variations. It is feature-agnostic and can be learned on raw input or extracted features such as HOG or VGG-Face. The figure has been taken from the published manuscript [122].

formulated as:

$$\min_{\mathbf{T},\mathbf{Z_1},\mathbf{Z_2}} \|\mathbf{TX_1} - \mathbf{Z_1}\|_F^2 + \|\mathbf{TX_2} - \mathbf{Z_2}\|_F^2 + \lambda\big(\epsilon\,\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\big) + \mu\,\|\mathbf{Z_2} - \mathbf{Z_1}\|_F^2 \tag{3.3}$$

where, the first two terms learn the shared transform and features, followed by the regularizers to prevent a degenerate solution, and a term to reduce the intra-class variations. The final term promotes features which minimize the difference between the representations of the same class, cross-domain pairs. While C-model handles only the intra-class similarity, cross-domain matching also suffers from the intertwined problem of low inter-class variations as well. In order to incorporate both intra-class and inter-class variations, building on the C-model, the *Divergent Model (D-model)* model is proposed.

Divergent Model (D-model): As shown in Figure 3-3, D-model utilizes the class labels at the time of feature learning, such that the samples of same class have similar representations, and representations of samples of different classes are far apart. With reference to Equation 3.2, $\mathbf{X_1}$ and $\mathbf{X_2}$ refer to cross-domain image pairs, and $\mathbf{y}$ contains the label corresponding to each pair specifying whether they belong to the same class (0) or different classes (1). In case of sketch to digital face image matching, the i^{th} sample of $\mathbf{X_1}$ and $\mathbf{X_2}$ may correspond to the sketch and digital

image of class-A with the i^{th} element of $\mathbf{y}$ as 0. On the other hand, another j^{th} sample of $\mathbf{X_1}$ and $\mathbf{X_2}$ may correspond to the sketch and digital image of class-A and class-B, respectively, with the j^{th} element of $\mathbf{y}$ as 1. This results in cross-domain pairs belonging to the same or different class. For a set of training pairs $\mathbf{X_1}$ and $\mathbf{X_2}$ belonging to two domains, with the label $\mathbf{y}$ (0 for same or 1 for different), $\mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$ represents a distance function which introduces discriminability during the feature learning process by handling the inter-class and intra-class variations. In this research, $\mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$ has been conceptualized as the contrastive loss [63], expressed as:

$$\mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y}) = (1 - \mathbf{y})\frac{1}{2}\left(\mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})^T\right) \; + \; (\mathbf{y})\frac{1}{2}\, max\big(0, (m - \mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})^T)\big) \tag{3.4}$$

where, $\mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})$ is a distance function applied on the learned representations $\mathbf{Z_1}$ and $\mathbf{Z_2}$ to model the relationship between the two, and m is the margin value. $\mathbf{y}$ is a $1 \times n$ vector containing the class labels, and the transpose operator ensures compatibility between the vectors. The margin ensures that only those inter-class pairs whose distance is less than the margin contribute to the loss function. The distance function $\mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})$ is modeled as the Euclidean distance between the representations $\mathbf{Z_1}$ and $\mathbf{Z_2}$. The above loss minimizes the Euclidean distance between cross-domain intra-class pairs, and focuses on those inter-class pairs which have a distance less than the margin m. After incorporating the contrastive loss, the proposed D-model is formulated as:

$$\begin{aligned} &\min_{\mathbf{T}, \mathbf{Z_1}, \mathbf{Z_2}} \|\mathbf{T}\mathbf{X_1} - \mathbf{Z_1}\|_F^2 + \; \|\mathbf{T}\mathbf{X_2} - \mathbf{Z_2}\|_F^2 \; + \; \lambda\big(\epsilon \, \|\mathbf{T}\|_F^2 - \; log \det \mathbf{T}\big) \\ &+ \mu\Big((1 - \mathbf{y})\frac{1}{2}\left(\mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})^T\right) \; + \; (\mathbf{y})\frac{1}{2}\, max\big(0, (m - \mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})^T)\big)\Big) \end{aligned} \tag{3.5}$$

where, similar to C-model, the first two terms learn the shared transform $\mathbf{T}$, and the representations $\mathbf{Z_1}, \mathbf{Z_2}$ for the two domains. The next two terms are the regularizers to prevent a degenerate solution, and the final two terms correspond to the distance loss applied on the representation. The proposed D-model thus learns a shared transform for cross-domain data. Representations are learned such that the intra-class variations across domains are minimized, and the inter-class separability is maximized. One of the major highlights of the proposed algorithm is that it is feature-agnostic, i.e. they can be applied to raw input or other features extracted from the images.

3.3.3 Optimization of C-model and D-model

Both C-model and D-model are optimized using the alternating minimization technique, which iteratively optimizes over all the variables. Each variable ($\mathbf{T}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$) is optimized in an alternate manner, while keeping the other variables constant. The following three steps are performed in an iterative manner for learning the C-model:

Update for $\mathbf{T}$:

$$\begin{aligned} &\min_{\mathbf{T}} \|\mathbf{T}\mathbf{X_1} - \mathbf{Z_1}\|_F^2 + \|\mathbf{T}\mathbf{X_2} - \mathbf{Z_2}\|_F^2 + \lambda\big(\epsilon \|\mathbf{T}\|_F^2 - log\det \mathbf{T}\big) \\ &\equiv \min_{\mathbf{T}} \left\|\mathbf{T}\begin{pmatrix}\mathbf{X_1}\\ \mathbf{X_2}\end{pmatrix} - \begin{pmatrix}\mathbf{Z_1}\\ \mathbf{Z_2}\end{pmatrix}\right\|_F^2 + \lambda\big(\epsilon \|\mathbf{T}\|_F^2 - log\det \mathbf{T}\big) \end{aligned} \tag{3.6}$$

The above equation is in the form of a standard transform learning model (Equation 2.5) and thus can be solved using the three step approach[148] as follows:

$$\mathbf{X}\mathbf{X}^\mathbf{T} + \lambda\epsilon\mathbf{I} = \mathbf{L}\mathbf{L}^\mathbf{T};\ \mathbf{L}^{-1}\mathbf{X}\mathbf{Z}^\mathbf{T} = \mathbf{U}\mathbf{S}\mathbf{V}^\mathbf{T} \tag{3.7}$$

$$\mathbf{T} = \mathbf{0.5V}\left(\mathbf{S} + (\mathbf{S^2} + \mathbf{2}\lambda\mathbf{I})^{\frac{1}{2}}\right)\mathbf{U}^\mathbf{T}\mathbf{L}^{-1} \tag{3.8}$$

The first step corresponds to a Cholesky decomposition on the input data, which is followed by performing a full Singular Value Decomposition. Finally, Equation 3.8 gives the update step for the transform matrix $\mathbf{T}$. In-depth proof of convergence and analysis of the above solution can be found in [148].

Update for $\mathbf{Z_1}$:

$$\min_{\mathbf{Z_1}} \|\mathbf{T}\mathbf{X_1} - \mathbf{Z_1}\|_F^2 + \mu\|\mathbf{Z_2} - \mathbf{Z_1}\|_F^2 \equiv \min_{\mathbf{Z_1}} \left\|\begin{pmatrix}\mathbf{T}\mathbf{X_1}\\ \sqrt{\mu}\mathbf{Z_2}\end{pmatrix} - \begin{pmatrix}\mathbf{I}\\ \sqrt{\mu}\,\mathbf{I}\end{pmatrix}\mathbf{Z_1}\right\|_F^2 \tag{3.9}$$

Update for $\mathbf{Z_2}$:

$$\min_{\mathbf{Z_2}} \|\mathbf{T}\mathbf{X_2} - \mathbf{Z_2}\|_F^2 + \mu\|\mathbf{Z_2} - \mathbf{Z_1}\|_F^2 \equiv \min_{\mathbf{Z_2}} \left\|\begin{pmatrix}\mathbf{T}\mathbf{X_2}\\ \sqrt{\mu}\mathbf{Z_1}\end{pmatrix} - \begin{pmatrix}\mathbf{I}\\ \sqrt{\mu}\,\mathbf{I}\end{pmatrix}\mathbf{Z_2}\right\|_F^2 \tag{3.10}$$

Equations 3.9 and 3.10 are least square problems having closed form solutions. The above three

Algorithm 1: Step wise optimization technique for training the D-model.

Input : $\mathbf{X_1}, \mathbf{X_2}, \mathbf{y}$
Output: $\mathbf{T}, \mathbf{Z_1}, \mathbf{Z_2}$

while *MaxIter* **do**

$$\mathbf{T} \leftarrow \arg\min_{\mathbf{T}} \|\mathbf{TX_1} - \mathbf{Z_1}\|_F^2 + \|\mathbf{TX_2} - \mathbf{Z_2}\|_F^2 + \lambda\left(\epsilon \|\mathbf{T}\|_F^2 - log \det \mathbf{T}\right)$$

$$\mathbf{Z_1} \leftarrow \arg\min_{\mathbf{Z_1}} \|\mathbf{TX_1} - \mathbf{Z_1}\|_F^2 + \mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$$

$$\mathbf{Z_2} \leftarrow \arg\min_{\mathbf{Z_2}} \|\mathbf{TX_2} - \mathbf{Z_2}\|_F^2 + \mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$$

end

updates are repeated iteratively until convergence or maximum iterations. Similarly, Algorithm 1 presents the step-wise approach for optimizing D-model.

3.4 Experimental Setup

To demonstrate the applicability of the proposed approach across different sketch to digital image matching applications, experiments are performed with seven datasets pertaining to three case studies: (i) caricature to face image matching, (ii) sketch-based object image retrieval, and (iii) face sketch to digital image matching. Most of the experiments are performed under the real world scenario of limited data availability. Details regarding the datasets and protocols are given below.

3.4.1 Datasets and Protocol

The proposed C-model and D-model have been evaluated on seven datasets (as shown in Figure 3-4) for three different cross-domain matching applications. Details regarding each dataset are given in Table 3.1. Unless explicitly mentioned, pre-defined protocols provided in the respective publication of each dataset have been followed for generating the training and testing splits. Each protocol results in disjoint training and testing partitions.

Case Study-1: Caricature Face Recognition: With the advent of social media and availability of multiple communication platforms, the use of caricatures has increased tremendously[2]. This

[2]https://tinyurl.com/y8z2kje6

(a) IIIT CFW (b) Caricature Face

(c) Shoe (d) Chair

(e) IIIT-D Forensic (f) CUFS (g) CUFSF

Figure 3-4: Sample images from the datasets used in experiments. The first row of each dataset contains the digital images and the second row contains images of the other domain.

creates a need for performing caricature to digital face image matching[3], applicable in several scenarios including image retrieval and social media profile linking. Results have been demonstrated on two publicly available datasets:

- IIIT Cartoon Faces in the Wild (IIIT-CFW) dataset [115] contains 8,298 caricatures and 1,000 face images of 100 public figures. Pre-defined training and testing splits have been used for performing the recognition experiment, where data corresponding to 42 identities is used for training, while the remaining 58 subjects form the test set.

- Caricature dataset [84] contains paired caricature and digital face images of 207 subjects, such that each subject has a single face image and a caricature image. The dataset contains both hand-drawn and digital caricature images collected from the Internet, thereby making

[3]Caricature face recognition refers to matching a caricature face image with a gallery of digital face images.

Table 3.1: Dataset details for the three case studies of sketch to digital image matching. 'Extnd. Gallery' refers to the number of images in the extended gallery for that dataset. Domain-A refers to digital photographs, while Domain-B refers to the sketch based images.

Case-Study	**Dataset**	**Number of Images in**	
		Domain-A	**Domain-B**
Caricature Face Recognition	Caricature DB [84]	207	207
	IIIT CFW DB [115]	1,000	8,298
Sketch Based Image Retrieval	Chair DB [203]	297	297
	Shoe DB [203]	419	419
Sketch Face Recognition	IIIT-D Forensic Sketch DB [18]	190 + 7,125 Extnd. Gallery	190
	CUFS DB [185]	188	188
	CUFSF DB[209]	1,194	1,194

the problem further challenging. On this dataset, images pertaining to 138 subjects are used for training, while the remaining 69 subjects form the test set.

Case Study-2: Sketch Based Image Retrieval (SBIR): SBIR refers to the task of retrieving digital images corresponding to a probe sketch image. SBIR has received dedicated attention for its applicability in scenarios of online shopping, similar object retrieval, and image based search. Experiments have been performed on two datasets:

- Chair dataset [203] contains 297 paired sketch and photo images. Experiments are performed using the pre-defined protocol, where 200 pairs are used for training, while the remaining form the test set.
- Shoe dataset [203] contains 419 pairs of sketch and photo images. Experiments are performed using the pre-defined protocol, where 304 pairs form the training partition, and the remaining form the test set.

Case Study-3: Sketch Face Recognition: Sketch face recognition has been a long standing problem which involves matching a sketch face image with its corresponding digital face image. Due to its applicability in forensic scenarios, it has also received substantial attention; however, matching forensic sketch images with digital face images remains a challenging and unsolved task. The reasons include larger intra-class variations and availability of limited training data. Results have been demonstrated on three publicly available datasets - two for viewed sketch face recognition and one for forensic sketch recognition. Details of each database are given as follows:

- IIIT-D Forensic Sketch dataset [18] has been used for evaluating the proposed model for forensic sketch face recognition. It contains 190 forensic sketches with corresponding digital face images. The sketches have been created based on the description provided by an eye-witness to the artist in real-world crime scene scenarios. Consistent with existing literature [123], and simulating the real-world scenarios, matching has been performed against an extended gallery of 7,265 digital face images, including 190 digital face images from the IIIT-D Forensic Sketch dataset. Due to the lack of a training set for the IIIT-D Forensic Sketch dataset, cross-dataset experiment is performed, where IIIT-D Semi-Forensic dataset [18] has been used for training. The IIIT-D Semi-Forensic dataset contains 140 sketch-digital face image pairs which simulate semi-forensic scenarios. It is ensured that the training and testing sets are subject disjoint.

- CUHK Face Sketch (CUFS) dataset [185] contains viewed sketches hand-drawn by an artist for a frontal image captured under normal lighting conditions with a neutral expression. Consistent with the existing protocol in literature [179], experiments have been performed on the CUHK student dataset, containing 188 sketch-photo pairs. 100 image pairs are used for training, while the test set contains the remaining 88 image pairs.

- CUHK Face Sketch FERET (CUFSF) dataset [209] contains 1,194 images from the FERET dataset [140] and their corresponding shape exaggerated hand-drawn sketches, created while viewing the image with lighting variations. Consistent with literature [179], 297 pairs are used for testing and the rest form the train set.

3.4.2 Experimental Details

Since a key highlight of the proposed approach is its independence to the input feature, experiments are performed with four different inputs: raw pixel values, Local Binary Patterns (LBP) [126], Histogram of Oriented Gradients (HOG) [37], and a pre-trained deep learning based feature extractor. For face recognition, the VGG-Face model [133] is the pre-trained deep learning model, while ResNet-152 [65] has been used for experiments on object images. VGG-Face is a pre-trained CNN model on the large-scale VGG-Face dataset, and ResNet-152 is pre-trained on the ImageNet dataset [90]. In the literature, both the models have demonstrated state-of-the-art classification

performance for face and object recognition, respectively. Comparison has also been performed with state-of-the-art algorithms for each case-study.

3.4.3 Implementation Details

For training the C-model and D-model, same-class pairs are generated by combining each sample of domain A with all the samples of the same class in domain B. Same number of different-class pairs are generated by combining samples of domain A with randomly chosen samples of domain B from a different class. The proposed models are trained on the input data for 50 epochs. The margin parameter (m) for D-model has been set based on grid search. Classification is performed on the learned features using Euclidean distance. Data augmentation is performed by flipping across the y-axis and performing illumination variations. The dimensions of the existing models used for comparison like DeepTransformer [191] are the same as proposed in the paper. The neural network used with DeepTransformer is of dimension $[k/2, k/4]$. The proposed models have been implemented in the Matlab R2018a environment. VLFeat and MatConvNet have been used for the extraction of LBP, HOG, and deep learning features. The experiments were performed on a on a workstation with 64GB RAM and one Nvidia K40 GPU. Moreover, we observed that the time taken to perform matching is comparable with classic object and face recognition matching algorithms. The proposed approach takes less than 1 ms to perform a match for a pair of images.

3.5 Results and Analysis

Experiments have been performed by using (i) the features as it is (referred to as Original), (ii) vanilla transform learning [146] (TL only), where transforms are learned for each domain independently, (iii) Shared Transform Learning (STL), and the proposed (iv) C-model and (v) D-model (Table 3.2). As mentioned previously, results are demonstrated with different input features, i.e., (i) raw pixels, (ii) LBP, (iii) HOG, and (iv) deep learning based feature extractor. Comparison has also been performed with the other cross-domain matching techniques and state-of-the-art results reported on each dataset (Tables 3.3-3.6). Consistent with the existing protocols, for all the datasets, results are reported in the form of accuracy at top-k, where $k = 10$, i.e. percentage of test images whose corresponding true match was retrieved in the top 10 ranks. For some sketch

Table 3.2: Rank-10 matching accuracy (%) for caricature face recognition and sketch-based image retrieval. Accuracies have been reported after matching with the original features, transform learning (TL only) features, shared transform learning (STL) features, contractive model (C-model), and divergent model (D-model). The proposed models demonstrate improved performance across different input features.

Input	**Original**	**TL Only**	**STL**	**C-model**	**D-model**
Caricature Face Recognition					
Caricature Face Dataset					
Pixels	19.56	23.19	22.46	27.53	**30.43**
LBP	18.84	18.84	20.29	26.81	**26.81**
HOG	34.05	28.26	31.16	39.13	**39.85**
VGG-Face	44.93	61.59	68.11	69.57	**78.98**
IIIT CFW Dataset					
Pixels	22.44	23.47	23.49	24.53	**28.16**
LBP	26.71	25.27	25.89	33.74	**35.20**
HOG	32.70	29.89	29.19	34.17	**36.47**
VGG-Face	59.26	79.16	77.29	79.98	**86.05**
Sketch-based Object Image Retrieval					
Shoe Dataset					
Pixels	8.70	9.57	9.57	13.91	**18.26**
LBP	37.39	39.13	30.43	40.00	**52.17**
HOG	80.00	80.00	79.13	80.00	**86.96**
ResNet-152	39.13	46.09	26.96	50.43	**55.65**
Chair Dataset					
Pixels	14.43	20.62	14.43	25.77	**45.36**
LBP	61.86	63.92	55.67	77.32	**83.51**
HOG	94.85	93.81	92.78	95.88	**96.91**
ResNet-152	85.57	86.60	72.16	**94.85**	93.81

face recognition experiments, $k = 1$, i.e. rank-1 accuracy is reported. The following subsections present results pertaining to each case study, and analysis of the proposed models.

3.5.1 Case Study Specific Analysis

Tables 3.2 to 3.6 present the results obtained for the different sketch to digital image experiments. Analysis corresponding to each case-study is provided below:

Case Study-1: Caricature Face Recognition: Table 3.2 presents the rank-10 or top-10 accuracy of the proposed models with different input features. The proposed D-model+VGG-Face features

Table 3.3: Rank-10 matching accuracy (%) comparing the proposed D-model with other algorithms. Accuracies are reported on the Caricature Face Dataset (CF), and the IIIT CFW dataset. The proposed D-model and DeepTransformer models are trained with VGG-Face features.

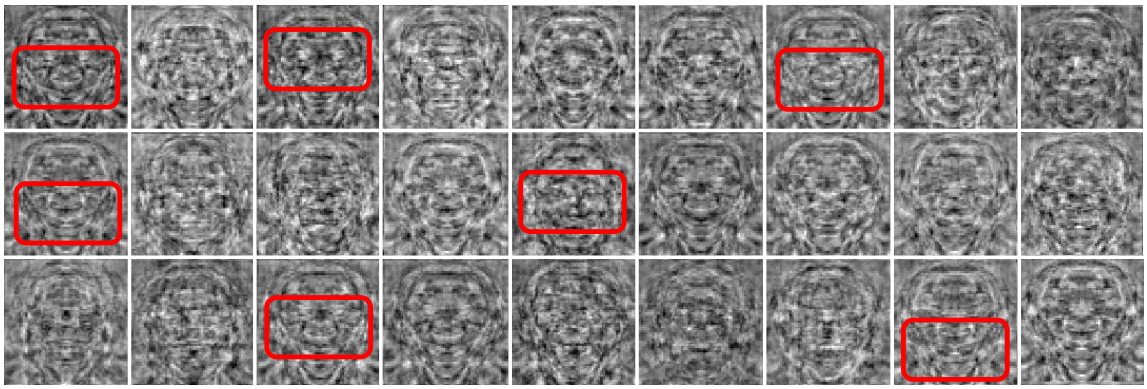
Algorithm	**CF**	**IIIT CFW**
COTS [4]	0	8.06
MMSS [175]	39.13	38.13
GSMFL [97]	47.10	41.01
Pixel + Neural Network	23.19	24.43
VGG-Face + DeepTransformer [123] + NNet	31.15	31.86
VGG-Face + Proposed D-model	**78.98**	**86.05**

performs the best by achieving 78.98% and 86.05% on the Caricature Face dataset and the IIIT-CFW dataset, respectively (Table 3.2). This can primarily be attributed to the fact that VGG-Face is a learning based model which has specifically been trained for faces, and thus demonstrates the best performance. Table 3.3 compares the performance of the proposed model with other recently proposed cross-domain algorithms, a Commercial off-the-shelf system (COTS) [4], and a neural network of dimension $[\frac{k}{2}\,\frac{k}{4}]$ trained on raw pixels. It can be observed that VGG-Face+D-model outperforms other cross-domain matching techniques by at least 45% on both datasets.

In literature, the only comparative accuracy on the Caricature Face Dataset is by Klare *et al.* [84], where the authors achieve a rank-10 accuracy of 74.8%. However, they performed experiments with 197 subjects, whereas the released dataset contains 207 subjects, which has been used in this study[4]. The proposed D-model achieves an improvement of almost 5% on this dataset. There do not exist any reported results on the IIIT-CFW dataset. Further, the proposed model outperforms the existing transform learning based cross-domain matching model, DeepTransformer [123], which demonstrates the benefit of encoding the class variations across domains during feature learning, and learning a shared transform as opposed to multiple transforms.

Figure 3-5(a) presents sample learned weights of the transform matrix on raw pixels. Outline of faces and caricatures, along with some salient features can be observed. For instance, some of the weights are characterized by exaggerated jaw lines and nose boundaries, while some contain balanced facial features. Upon analyzing the mis-classified samples for caricature face recognition (Figure 3-5(b)), we observed that most of the images were of varying pose. Exaggeration of facial

[4]For consistency with Klare *et al.* [84], one-third and two-third subjects are used for testing and training, respectively.



(a) Weights Visualization

(a) Subject-A (b) Subject-B (c) Subject-C

(b) Sample Mis-Classifications

Figure 3-5: (a) Visualizations of sample weights learned by the proposed D-model on the Caricature Face dataset, for pixel input. It can be observed that the model learns different components of faces and caricatures - both exaggerated and balanced. (b) Sample images mis-classified by the D-model.

features, extreme pose variations, and different artistic techniques renders the problem further challenging.

We have also performed additional experiments on the IIIT-CFW dataset for two recent few shot protocols provided by Zheng *et al.* [212]. For the task of Few-Shot Photo to Caricature Recognition, we obtain an accuracy of 86.73% which is 1.2% higher than what is reported by Zheng *et al.* [212]. Additionally, on the second protocol of Few-Shot Photo to Caricature Recognition, we obtain 97.64%, as opposed to 93.7% reported by authors. The results demonstrate the effectiveness of the proposed approach on benchmark datasets.

Case Study-2: Sketch based Image Retrieval (SBIR): As observed in Table 3.2, the proposed D-model with HOG features performs best by achieving 86.96% and 96.91% on the Shoe and Chair datasets, respectively. One of the key reasons for this behavior is that HOG features encode the gradient information, which provide important distinct information about sketches. Table 3.4 compares the matching accuracy of the proposed D-model with other techniques for sketch based object image retrieval. Owing to the fixed protocol of the dataset, results have directly been taken

Table 3.4: Rank-10 accuracy (%) of the proposed D-model and other algorithms on the Shoe and Chair datasets.

Algorithm	**Shoe DB**	**Chair DB**
BoW-HOG [203]	67.83	67.01
Dense-HOG [203]	65.22	93.81
ISN Deep [203]	62.61	82.47
Triplet model [203]	**87.83**	**97.94**
Song *et al.* [163]	**91.30**	**98.97**
Song *et al.* [164]	**94.78**	95.88
HOG + DeepTransformer [123] + NNet	53.91	91.75
HOG + Proposed D-model	**86.96**	**96.91**

from the respective publications. It can be observed that the proposed D-model with HOG features is among the top performing models for the given datasets. It is interesting to note that the proposed model mis-classifies only three samples of the chair database, whereas the best performing model mis-classifies a single sample [163], and most of the recent models have been pre-trtained on the ImageNet photo-edgemap pairs[164].

For the application of sketch-based object image retrieval, we observe that the proposed technique is able to model shape and visual features of the images quite well. Figure 3-6(a) presents the rank-list of two samples which are *not* classified by the proposed HOG+D-model. In the first case, all the retrieved samples have the same shape as the query image, and a similar pattern of legs with wheels. Similarly, in the second case, the retrieved samples have the same shape, which demonstrates the utility of the proposed model for retrieving similar images as well. Figure 3-6(b) presents the rank-list obtained by HOG features and HOG+D-model features for a sample query of the Shoe dataset. HOG+D-model is able to retrieve the correct match at rank-2, which is not the case with HOG features based matching. We believe that the D-model further enforces discriminability in the HOG features, thus enhancing the classification performance.

Case Study-3: Sketch Face Recognition: On the IIIT-D Forensic Sketch dataset, the proposed D-model (with HOG features as input) provides the best matching performance. Table 3.5 compares its performance with the current state-of-the-art results. Since we follow the same benchmark protocol, results have directly been taken from Nagpal *et al.* [123]. As compared to state-of-the-art DeepTransformer with L-CSSE (deep learned feature), the proposed D-model with HOG demon-

(a) Chair Dataset

(b) Shoe Dataset

Figure 3-6: Sample retrieval results for the Chair and Shoe dataset with the proposed D-model with HOG features as input.

strates an improvement of around 15%, and an improvement of over 32% is observed with COTS at rank-10. Figure 3-7 presents the Cumulative Match Characteristics (CMC) curves obtained on the forensic sketch dataset. At rank-50, an improvement of over 25% is observed with the proposed HOG+D-model, as compared to the current state-of-the-art. Figure 3-8 presents visualizations of sample weights learned by the transform learning model on the sketch images for pixel input. The model appears to encode the facial features, especially at the global level. For example, different weights appear to model varying face shape and hairstyles. Since face sketches tend to have more holistic information as compared to minute local features, modeling the global information enables better feature extraction, thereby facilitating improved classification performance.

Table 3.5: Rank-10 matching accuracy (%) of the proposed D-model and other algorithms on the IIIT-D Forensic Sketch dataset with a large-scale gallery.

Algorithm	**Forensic DB**
COTS [4]	2.63
MCWLD [18]	14.71
GSMFL [97]	13.46
L-CSSE + DeepTransformer [123] + NNet	19.44
HOG + Proposed D-model	**35.26**

Table 3.6: Rank-1 accuracy (%) of the proposed D-model and other comparative algorithms on the CUFS and CUFSF dataset.

Algorithm	**CUFS**	**CUFSF**
Pix2Pix [73]	100.0	37.0
CycleGAN [215]	99.0	25.0
DualGAN [201]	100.0	35.0
Multi-Adversarial Networks ($PS^2 - GAN$) [179]	100.0	47.0
Semi-coupled Dictionary Learning [183]	95.2	-
Multi-Paced Dictionary [197]	98.4	-
Generalized Coupled Dictionary Learning [107]	98.0	-
Locality-Constrained Joint Dictionary + Residual Learning [75]	98.2	-
Multi-view Domain Translation [137]	98.1	-
HOG + Proposed D-model	**100.0**	**67.3**

Table 3.6 presents the results on the two viewed sketch face datasets: CUFS [185] and CUFSF [209]. The proposed HOG+D-model achieves 100% and 67.3% on the CUFS and CUFSF dataset, respectively. Comparison has been drawn with existing Generative Adversarial Network (GAN) based techniques such as Pix2Pix [73], DualGAN [201], and CycleGAN [215] (owing to the same protocol, results have directly been taken from Wang *et al.* [179]). The proposed D-model with HOG features outperforms GAN based image synthesis techniques on both the datasets. Comparison has also been performed with existing dictionary learning based techniques (owing to the same protocol, results have directly been taken from their respective publications, and Mandal and Biswas [107]), where the proposed model demonstrates an improvement of atleast 1.5% on the CUFS dataset. The results showcase the efficacy of the D-model for sketch to photo face recognition.

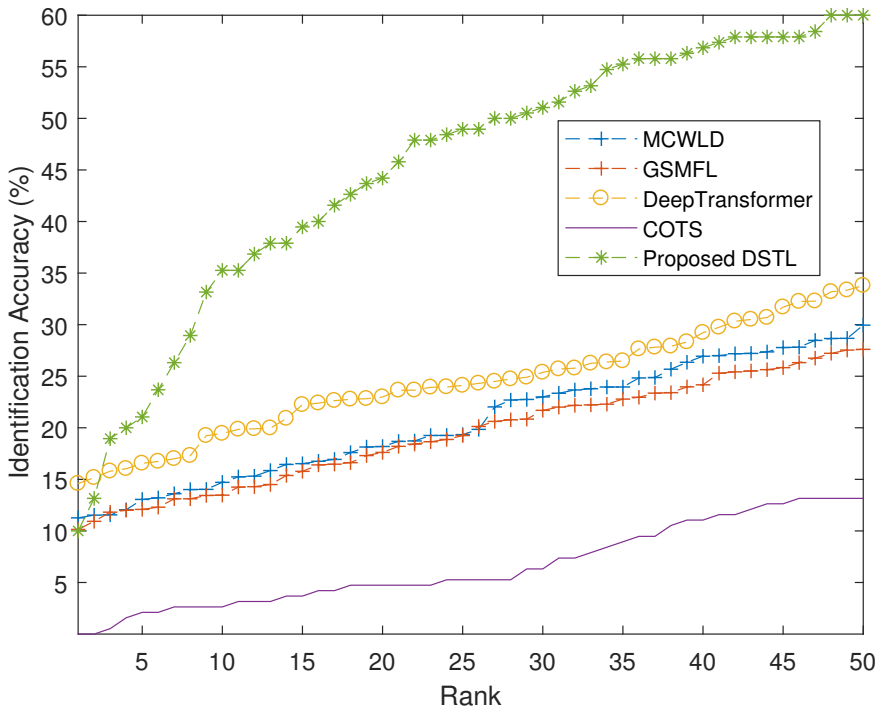


Figure 3-7: Cumulative Match Characteristics (CMC) curves on the IIIT-D Forensic Sketch database. The proposed DSTL algorithm (D-model) outperforms existing methods to report superior performance.

3.5.2 Overall Analysis of the Proposed Models

The proposed C-model and D-model have also been analyzed in terms of the effect of input feature, ablation study, and number of trainable parameters:

Effect of Input Features: The proposed models are evaluated with different input features, both hand-crafted and deep learning based. From Table 3.2, it is observed that both hand-crafted and deep learning based features perform better than raw pixels, across all case-studies. This can be attributed to the challenging nature of the problem, which require informative features for matching across domains. Moreover, features which outperform others without any transform, attain the best results when given as input to the D-model as well. For example, VGG-Face features (original) yields the highest accuracy as compared to other features on the Caricature Face dataset and the IIIT CFW dataset. With the proposed D-model, it achieves state-of-the-art performance with an improvement of around 26-33%. Further, for shoe and chair databases, handcrafted features yield better results as compared to deep learning features.

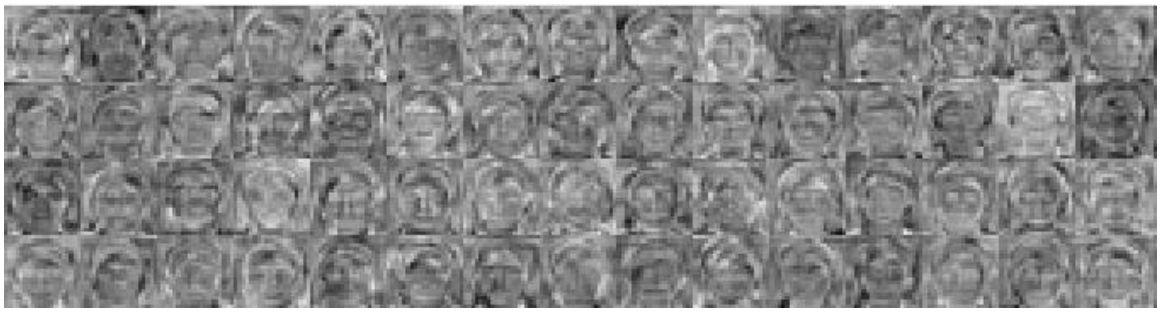
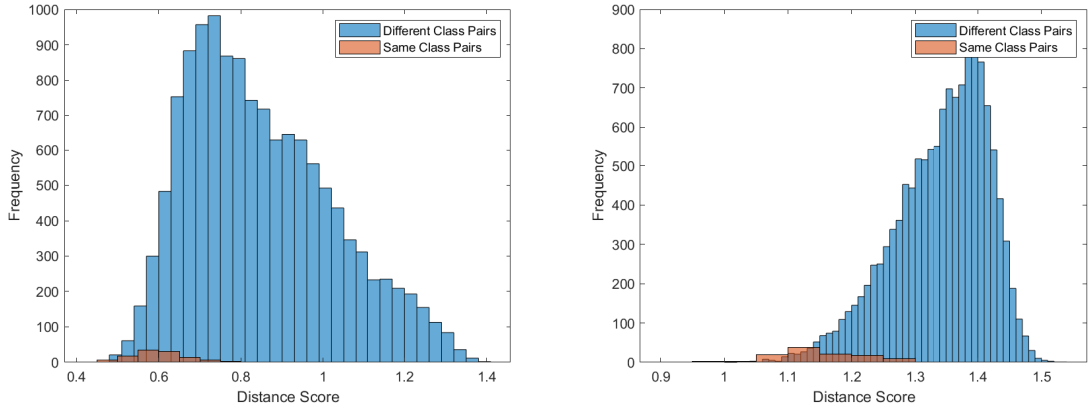


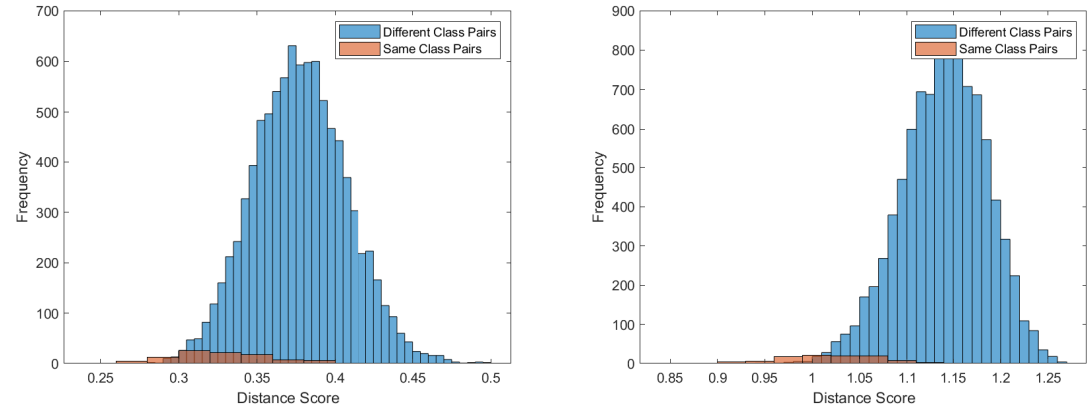
Figure 3-8: Visualizations of sample weights learned by the proposed D-model algorithm, for sketch face recognition on raw pixels.

Ablation Study: Table 3.2 also presents the ablation study performed on the proposed D-model. The table can be read from right to left to analyze the effect of each component. D-model embeds both intra-class and inter-class variations, while C-model focuses on reducing the intra-class variations only. The Shared Transform Learning (STL) model does not utilize the class labels, and only projects the input onto a common space. Further, Transform Learning only (TL only) eliminates the use of shared transform and uses two transforms for feature learning, without any class information. The first column, Original, uses Euclidean distance directly on the input for matching. Across different case-studies and input features, the proposed D-model achieves highest classification performance, as compared to other techniques. C-model, which focuses on reducing the intra-class variations only, demonstrates improved performance as compared to the original features, and unsupervised Transform Learning models. This motivates the usage of reducing intra-class variations while learning representations. Upon adding the term for modeling the inter-class separability as well (D-model), the improvement in accuracy is further pronounced. In most cases, across features, D-model outperforms C-model, which motivates the utility of the proposed model for different cross-domain matching tasks. The classification accuracy of TL only and STL models can be compared with C-model and D-model in order to further promote the inclusion of class information in the proposed models. An improvement of 4-31% is observed from STL to D-model, which can be attributed to the discriminative supervision constraints. It is interesting to note that while TL only improves the classification performance as compared to the Original features in almost all scenarios, similar improvement is not observed for Shared Transform Learning. This implies that projecting cross-domain data onto a common space, without additional supervision



(a) Shared Transform Learning (b) Divergent Model (D-model)

(A) Shoe Dataset



(a) Shared Transform Learning (b) Divergent Model (D-model)

(B) Chair Dataset

Figure 3-9: Score distributions obtained after (a) Shared Transform Learning and (b) Divergent Model (D-model) on two datasets for sketch based image retrieval. The separation of same and different class scores promote inclusion of discriminative class-specific terms in the proposed model (best viewed in color).

constraints may not enhance classification. The benefit of incorporating discriminative supervision constraints can also be observed from Figure 3-9. Modeling the inter-class and intra-class variations provides a separation between the same-class and different-class pairs, thus enhancing the classification performance. Further, D-model pushes different-class pairs to have a higher distance score as compared to the same-class pairs.

Number of Parameters: For a vector input of dimension $n \times 1$, and a square shared transform

of $n \times n$, a representation of dimension $n \times 1$ is obtained. Here, the proposed C-model and D-models only learn $n \times n$ parameters for the single shared transform. On the other hand, existing cross-domain Transform Leaning model, DeepTransformer [123], requires learning $3 \times n \times n$ parameters. This strengthens the usage of the proposed model, both in terms of reduced parameters and improved performance, probably for problems with limited training data.

3.6 Additional Case Study: Skull Recognition

Similar to the sketch face recognition, skull identification is a challenging problem and of vital importance in forensic and law enforcement scenarios. Given the nature of the problem and the limited availability of data, skull to face matching is another use-case for performing facial recognition with scarce data. Human skull identification refers to matching a skull image with gallery face images. Unidentified skulls often result in unrest with the law enforcement officials and family of the victim. For instance, in 2014, skeletal remains (including a skull) were found in North Creek, just above Virginia [152]. Forensic experts were able to predict an age-range, gender, and ethnicity of the skull, however, the identity remains unknown. Efforts were made to investigate missing person reports around that area, and other unsolved cases. Retired police chief, Kurt Wright, who worked on the case, mentioned "This has always just kind of nagged at me for some reason, and I wanted some closure to it" [152]. In 2017, the case was again publicized in news to look for fresh leads; however, the case remains unsolved.

Skull identification is primarily performed by means of a manual reconstruction approach. That is, given a skull image, a facial reconstruction is created which is then circulated via digital or print media, or matched with a given gallery of digital face images. The techniques for facial reconstruction can broadly be divided into three categories: (i) 2-D reconstruction, (ii) 3-D reconstruction, and (iii) superimposition. 2-D reconstruction techniques construct a 2-D face image from a given skull image, whereas 3-D reconstruction constructs a 3-D face, using either Computer Graphics or via clay or sculptures. Superimposition refers to the task of superimposing a digital face image over a skull, and analyzing its alignment accuracy. This technique is generally applied when the forensic experts have a small subset of possible identities to be verified against a given skull. Traditionally, the successful application of the above techniques require at least a couple of days to

a couple of weeks, along with collaboration between forensic artists and anthropologists. In this research, we propose to automate the process of skull identification. If a completely intact skull is found, it may enable the law enforcement agencies to determine its identity without performing the facial reconstruction. With the availability of a skull recognition algorithm, the search for the identity of a skull can extend beyond the immediate location in a fast, effective, and automated manner.

Skull identification can be viewed as a cross-domain matching problem, where one domain corresponds to skull images, while the second contains digital face images. It can be observed that the information content of digital face images varies significantly from that of the skull image.Cross domain matching has received significant attention in literature, in terms of low resolution face matching, or sketch to face matching, or visible to near infrared (NIR) face matching [124, 129]. However, the task of automated skull recognition has received limited attention [13, 38]. Most of the existing algorithms focus on the task of facial reconstruction [72] by means of superimposition, that is, scenarios having the availability of a small subset of possible identities for a given skull. Besides reconstruction, researchers have also focused on automating the task of matching 3-D CT head scans with a gallery of digital images [46, 171]. While this is a viable option to perform matching, it incurs the overhead charge of conducting a CT head scan, and the inherent assumption that the skull found can be used for scanning. A model for automated skull recognition without the need of reconstructions or scans has been proposed by Nagpal *et al.* [118]. The authors also proposed a publicly available IdentifyMe dataset consisting of 464 skull images, along with Semi-supervised transform learning (SS-TL) and Unsupervised transform learning (US-TL) models for skull to digital face image matching.

Existing algorithms for skull recognition rely either on the availability of CT head scans or require learning a large number of parameters. Given the availability of limited training samples, this might result in over-fitting and thus lower the identification performance on unseen probes. In order to address these limitations, proposed *Discriminative Shared Transform Learning* , which learns a single "shared transform" for both skulls and digital face images, while reducing the intra-class variations. Experimental results on the publicly available skull dataset, IdentifyMe [118] demonstrate the efficacy of the proposed model for the task of skull to face matching.

(a) The first row corresponds to the digital face images, while the second contains their respective skull images.

(b) Unlabeled supplementary skull images.

Figure 3-10: Sample images from the proposed IdentifyMe - Skull and Face Image Database. Thess images have been taken from Nagpal et. al [118].

3.6.1 Dataset and Protocol

There exists only one publicly available dataset for performing skull to digital face image matching, the IdentifyMe dataset [118]. The dataset contains 464 skull images, inclusive of 35 mated skull images and their corresponding face images. Figure 3-10 presents sample images of the IdentifyMe dataset. The authors have also provided two protocols and baseline results for evaluating the performance of any algorithm. The two protocols are as follows:

- **Protocol-1**: This protocol evaluates the performance of a given algorithm on the 35 mated pairs of skull and digital face image. The results are reported with five fold cross validation.

- **Protocol-2 (extended gallery)**: This protocol attempts to mimic a real world scenario where the gallery consists of more subjects as compared to those present in the dataset. An extended gallery of 993 subjects is used for evaluation, such that for a given fold, the gallery contains 1000 subjects (993 extended and 7 from the IdentifyMe dataset) at test time. Similar to the previous protocol, five fold cross validation is performed for evaluation.

Table 3.7: Rank-1 identification accuracy (%) of the proposed DSTL model and other algorithms on the IdentifyMe skull dataset.

Algorithm	**Protocol-1**	**Protocol-2**
HOG [37]	34.2	25.7
HOG + US-TL [118]	37.1	34.2
HOG + SS-TL[118]	42.9	37.1
HOG + Proposed C-modelL[161]	51.4	42.9
HOG + Proposed D-model	**54.2**	**42.9**

Face images of IdentifyMe dataset are detected using Viola Jones face detector [174], followed by geometric alignment with the skull images. Data augmentation is performed by combinations of flipping across the y-axis and modifying the brightness or contrast of images. Features are extracted from the pre-processed images, which are then used as input for the proposed models.

3.6.2 Results

Table 3.7 presents the rank-1 identification accuracy of the proposed DSTL algorithm on the two protocols of IdentifyMe dataset. The proposed model achieves improved performance on both the protocols, with HOG input features. An improvement of over 20% is observed with the D-model+HOG model as compared to the accuracy reported only by the HOG features. Improvement is also observed with respect to the current state-of-the-art results. The results indicate the benefit of incorporating discriminative supervision into the transform learning model. While the proposed model achieves improved performance for the challenging task of skull identification, there still exists a large margin for further improvement.

3.7 Summary

Almost all machine learning problems suffer from the challenge of low inter-class and high intra-class variations. These challenges are further pronounced in sketch to digital image matching tasks, due to the added variability in the information content across domains, and the availability of limited labeled training data pairs. This research proposes a novel *Discriminative Shared Transform Learning* algorithm, which utilizes a shared transform to perform effective feature extraction for data belonging to two domains under supervised constraints. Learning a shared transform under

supervised constraints enables the model to learn domain invariant features useful for enhanced classification performance. Further, two models have been presented under the proposed algorithm: Contractive and Divergent Models (C-model and D-model). D-model focuses on learning domain invariant representations while modeling the inter-class and intra-class variations across domains. Experimental analysis demonstrates that modeling both inter-class and intra-class constraints improves the performance of existing features in most cases. A unique characteristic of the proposed algorithm is its feature agnostic behavior, i.e. given different input features (raw pixels, hand-crafted, or deep learning based), it results in improved performance for the given task. For example, an improvement of around 4-34% is observed for caricature face recognition. Such techniques provide the flexibility of utilizing task-specific features, to boost the existing performance. A thorough evaluation of the proposed models has been performed on three different case studies classified under sketch to digital image matching: (i) caricature face recognition, (ii) sketch based object retrieval, and (iii) sketch face recognition. Results are also demonstrated on an additional challenging case study for face recognition with scarce data- Skull recognition The results for all case-studies demonstrate the efficacy of incorporating (i) shared (ii) and discriminative (contractive and divergent) terms in the existing transform learning model, resulting in improved state-of-the-art performance. For example, improvement of around 15% is observed on the challenging IIIT-D forensic sketch dataset. Despite the improved performance achieved by the proposed models, there still exists a vast scope for improvement, especially for real world scenarios. This research has been published in the Pattern Recognition Journal, 2021, and IEEE Internation Joint Conference on Biometrics (IEEE IJCB), 2017, and all the figures have been taken from the published research papers [122], [118].

Chapter 4

In-group Bias in Deep Learning based Face Recognition Models due to Ethnicity and Age

4.1 Introduction

Bias is often defined as an *inclination or prejudice for or against one person or group, especially in a way considered to be unfair*[1]. Recently, bias has been reported in several facial analysis systems, which highlights the concerns with present day state-of-the-art deep learning based facial analysis models [144]. In 2018, American Civil Liberties Union (ACLU) performed a study on Amazon's publicly available facial recognition software to analyze the growing concerns regarding biased behavior of AI systems against people of color [6]. This one of the many incidents shedding light on the emerging *biased* behavior of Artificial Intelligence systems, and the need to develop *fair AI*. Earlier examples include the ProPublica study of algorithms being used by the US courts and parole boards to predict future behaviour of criminals [5]. It was observed that the model was biased against black defendants, and gave them a higher score as compared to their white counterparts. Multiple incidents suggesting the existence of bias demand dedicated research to ensure the development of AI systems for the *better*, and not for the *worse*.

[1] https://tinyurl.com/yuwhntt5; https://www.lexico.com/definition/bias

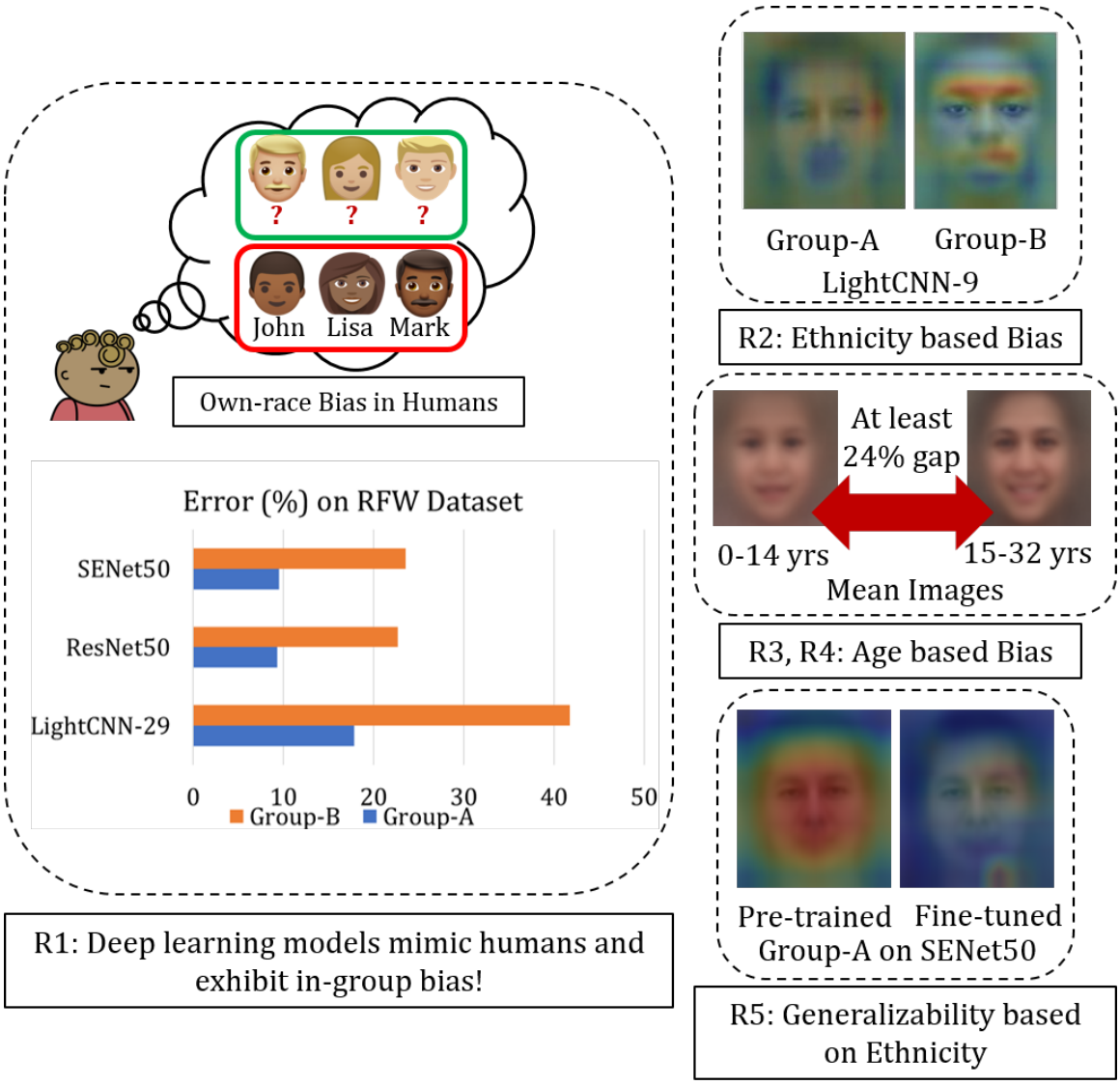


Figure 4-1: Summarizing the key results of this research demonstrating the biased nature of deep learning based face recognition models. R1: Own-race bias prevalent in humans exists in deep learning based face recognition models as well (error (%) on the RFW dataset). R2: Deep learning models trained on a single demographic group appear to focus on different discriminative regions for recognition. R3-R4: Pre-trained deep learning models demonstrate large gap in accuracy while recognizing face images of different age-groups. R5: Fine-tuning pre-trained models results in a shift of the area of focus, which might result in reduced generalizability.

This research focuses on understanding the hidden biases of *face recognition models*, in particular the *in-group bias* (Figure 4-1). In-group bias refers to the tendency to have an affinity towards members belonging to the same group as the individual. While bias in AI is a relatively new field, neuro-cognitive researchers have long established the presence of *in-group* bias in humans. We tend to recognize individuals of our demographic group rather easily, as compared to individuals of an *out-group* [125]. One such instance dates back to 1971, when five black men were acquitted for the murder of Khomas Revels [113]. Despite the lack of any forensic evidence, five eyewitnesses testified and identified them as offenders. However, during the third trial, private investigators located the actual culprits, who were later convicted. Interestingly, all the five men initially acquitted were black and all the eyewitnesses were white. Social Psychologist, Dr. William Haythorn identified the reason for mis-identifications to be the cross-racial identifications, because of which people of the other ethnicity look alike, resulting in biased outcomes. Several studies later, this phenomenon was termed as the *in-group effect* and in this case of ethnicity, it was also referred as *own race bias*, or the *other race* or *cross-race effect*.

Owing to the several recent incidents of biased predictions, it is essential to understand *if* and *where* bias is being encoded. Further, since most of these algorithms are based on deep learning, they provide limited insights into the decision making process. Therefore, understanding the models and their corresponding biases is of utmost importance, in order to prevent the inclusion of biases that existed in the past into our future. We attempt to analyze the biases being encoded, and explore whether the bias present in Deep learning based models exhibit in-group effect similar to humans. Such analysis can help us understand the phenomenon of bias in machines better, and therefore mitigate as well as prevent it.

4.1.1 Research Contributions

As shown in Figure 4-2, this research extends beyond the existing literature by analyzing deep learning based *face recognition models* for in-group effect. This is the first of its kind research which attempts to understand what deep networks are encoding, and whether they learn features similar to humans for different groups based on a given co-variate. Based on the observations of this research, a novel *bias index* has also been proposed for estimating the level of bias of a trained

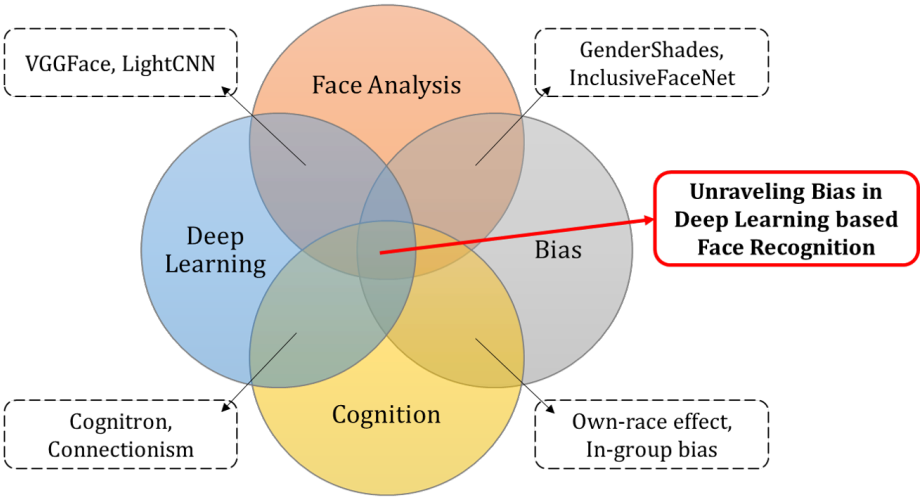


Figure 4-2: This research operates at the intersection of four key areas by analyzing *if* and *where* bias exists in deep learning based face recognition models (best viewed in colour).

facial recognition network.

Until now, face recognition algorithms focused on learning models with as much data as possible, without explicitly focusing on capturing the variations with respect to different sub-groups. Models were trained with the assumption that with the available large data, they will learn the sub-group variabilities in face images. In this research, we try to answer why limited exposure to certain groups is a challenge even when the models are trained with large databases. The in-group effect has been well studied in psychology, and this is the first work which simulates the same concept in machine learning. With the help of multiple experiments, we demonstrate that faces belonging to different groups follow different distributions and have different discriminative regions. Based on our findings we highlight the need to study each group and utilize representative data to learn unbiased models. It is our belief that coupled with the proposed bias index, insights into the functioning of deep networks can help develop fairer AI systems, capable of unbiased decisions. Since deep learning models appear to mimic the human behaviour of in-group/out-group effect, findings of this research can also enable better utilization of pre-trained models by applying suitable cognitive techniques (designed for humans) for de-biasing. Four deep learning networks (LightCNN-9 [192], LightCNN-29 [192], ResNet-50 [65], and SENet-50 [68]), either trained from scratch or pre-trained on large scale datasets containing about 10 million images are used for our findings. The significant contributions of this research are two-fold as follows:

1. Key results of this research inferred from the experiments and evaluation are as follows:

    - **Result-1:** Simulating the real world scenario of limited exposure to all groups, by training on data belonging to a specific group only, we observe that deep learning models mimic the human tendency of *in-group bias*. We study the tendency across two co-variates of ethnicity and age.
    - **Result-2:** Similar to the human behavior of face recognition, ethnicity-specific *regions of interest* are encoded in networks trained on a specific ethnicity.
    - **Result-3:** Extending upon the existing cognitive literature, different regions of interest are also observed for varying age-groups, suggesting *in-group effect with respect to age (own-age bias)* in deep learning networks.
    - **Result-4:** Networks pre-trained on large-scale data with varying distribution (eg. VGGFace2 [25] and MSCeleb-1M [62] datasets), demonstrate high generalization abilities, however, bias across age persists to be a major challenge.
    - **Result-5:** Fine-tuning a pre-trained network results in change of *region of interest*, often reducing its generalizability, reactivating the in-group effect.

2. **Bias Index:** Based on the derived observations, a novel bias index has been proposed for predicting a trained face recognition network's *level of bias*. The bias index models the level of localization of the regions of interest learned by a given model, thus providing an estimate of a learned network's bias for feature extraction.

4.2 Related Work

Prior to research in the Machine Learning domain, researchers have studied the effect of in-group bias in humans for several years [22]. This effect has been observed across various spectra ranging from hiring, financial markets, to elections [28, 74, 157]. It has also been studied via cognitive studies on humans with respect to facial recognition and bias [31].

With bias emerging as a challenge for face recognition systems, research efforts have been made to formally define *bias* and *fairness* [53], and to develop techniques for quantifying bias of

a machine learning model [59]. Limited research has also focused on developing unbiased and transparent machine learning algorithms [9, 27, 105, 120]. Research in this domain has primarily focused on understanding bias and fairness in models developed for textual data [20, 143], with limited attention to computer vision tasks [57, 66, 184]. Researchers have explored the presence of bias in commercial-off-the-shelf systems and deep learning models primarily for gender and ethnicity classification in face images [24, 39, 155]. Buolamwini and Gebru [24] categorized face images into four categories: darker males, darker females, lighter males, and lighter females. It was observed that, dark-skinned females were the most mis-classified group, thereby creating a need for fairer unbiased facial analysis algorithms. It is important to note that the authors analyzed gender classification using three commercial systems, without exploring the performance of deep learning models.

Parallelly, research efforts are also being made to develop mitigation strategies as well as bias invariant learning methods. Amini *et al.* [11] presented a de-biasing technique for face detection algorithms with respect to gender and ethnicity. The authors proposed learning the latent structure in the training data using variational autoencoders, followed by re-weighting the importance of the training data samples for the given task. Das *et al.* [39] proposed a Multi-Task Convolutional Neural Network (MTCNN) for joint gender, ethnicity, and age classification of face images. The proposed model demonstrates improved performance for the given tasks across multiple subgroups of gender, age, and ethnicity. Simultaneously, Alvi *et al.* [10] proposed a joint learning and unlearning (JLU) technique for eliminating bias from CNN networks. Experiments are performed for multiple tasks of age, gender, ethnicity, and pose classification of face images, independently, while the other covariates are used for inducing bias. Ryu *et al.* [155] proposed using transfer learning to develop attribute prediction models inclusive of different ethnicity and gender subgroups via a novel InclusiveFaceNet model. Recently, Wang *et al.* [181] presented a novel Racial Faces in the Wild dataset, containing equal distribution of face images from different ethnicities. The authors also proposed a novel domain adaptation technique for eliminating face recognition bias from deep learning networks. Recently, Robinson *et al.* [153] proposed the Balanced Faces in the Wild dataset containing data balanced across gender and ethnicity. The authors reported the accuracy variations across sub-groups and proposed learning group specific thresholds.

However, to the best of our knowledge, none of the existing techniques have analyzed deep

Table 4.1: List of 40 experiments performed to understand face recognition in deep learning models w.r.t. ethnicity and age bias. Results have been shown using three datasets for ethnicity: CMU Multi-PIE, MORPH, and RFW; and two datasets for age: Adience and CACD. MS-Celeb-1M, CWF: CASIA-WebFace, VF2: VGG-Face2; G-A: Group-A, G-B: Group-B; Age-1 (A-1): 0-14 years, Age-2 (A-2): 15-32 years, Age-3 (A-3): 33+ years.

Exp.#	Network	Pre-train	Train	Fine-tune	Test
Analysis of Bias due to Ethnicity					
1 / 2	LightCNN-9	-	G-A	-	G-A / G-B
3 / 4			G-B	-	G-A / G-B
5 / 6			G-A, G-B	-	G-A / G-B
7 / 8	ResNet-50	-	G-A	-	G-A / G-B
9 / 10			G-B	-	G-A / G-B
11 / 12		MSC (10M),	-	-	G-A / G-B
13-16			-	G-A / G-B	G-A / G-B
17 / 18	LightCNN-29	CWF (50K), MSC (10M)	-	-	G-A / G-B
19-22				G-A / G-B	G-A / G-B
23 / 24	SENet-50	VF2 (3M)	-	-	G-A / G-B
25-28			-	G-A / G-B	G-A / G-B
Analysis of Bias due to Age					
29	LightCNN-9	-	A-1	-	A-1
30			A-2	-	A-2
31			A-3	-	A-3
32-34	LightCNN-29	CWF (50K), MSC (10M)	-	-	A-1 / A-2 / A-3
35-37	ResNet-50	MSC (10M), VF2 (3M)	-	-	A-1 / A-2 / A-3
38-40	SENet-50		-	-	A-1 / A-2 / A-3

learning models to understand how or why bias is encoded, especially with respect to the in-group effect. This research is among the few research directions which explores and analyzes the in-group bias in deep learning models, specifically, if they encode bias and exhibit the in-group effect similar to humans. It also attempts to answer *if* and *where* bias is being encoded by visually analyzing various deep networks and their behaviour in different setups.

4.3 Methods and Hypotheses

In this chapter, we analyze the behavior of deep learning networks with respect to well established cognitive studies on humans, and identify the regions used by deep learning models to learn features for face recognition. Through our experiments, several deep-learning based face

recognition networks are analyzed to answer a key question: *Do deep learning systems encode in-group biases?* We delve deeper to observe if these regions are consistent for a given attribute (ethnicity/age-group), across different values, and how deep learning models behave when trained from scratch on selective data i.e. data highly biased towards specific subgroups. In this study, we attempt to answer the above mentioned key question by studying two attributes for face recognition - *ethnicity* and *age*. Thus, this chapter and experiments are structured in order to specifically answer the following questions:

- Does deep learning encode in-group bias w.r.t. ethnicity?
- Does deep learning encode in-group bias w.r.t. age?

Experimental Setup: Analysis is drawn from 40 experiments (Table 4.1) conducted across different attributes, networks, and training data on the following two setups:

- Networks trained-from-scratch: Deep learning models are used to simulate the human behavior of in-group effect, where an individual is acquainted with people of a specific group only (e.g. based on ethnicity). The trained models are used to analyze if the deep learning models exhibit the in-group effect by (i) focusing on different discriminative regions for classification (similar to humans) and (ii) and obtaining poor performance on faces from a different group. To this effect, models are trained on data pertaining to a single group only (ethnicity/age), to observe the regions being learned for face recognition; and on equal number of samples belonging to different groups. Additionally, models are also evaluated on alternate groups (different group from the training group) to understand what regions are being used in each case to perform classification. For example, LightCNN-9 trained on Group-A subjects is used for evaluation on Group-B subjects, to study the difference in processing.
- Pre-trained Networks: Pre-trained face recognition models have been used to understand whether training on large-scale data result in (i) near equal recognition accuracy across different groups, and (ii) reduced focus on specific facial regions during feature extraction and classification. The effect of fine-tuning is also studied to understand how the regions of interest evolve for different groups with pre-trained networks. Fine-tuning is performed on the entire network using the train set.

In both the setups, feature-level analysis is performed with respect to the learned feature maps, where the final convolution layer maps are interpolated and super-imposed on the input image, to obtain the region of interest for the given input and network. Class Activation Maps (CAMs) [214] are used to obtain the most discriminative regions of face images, focused upon by the CNNs. Results are also reported in terms of the verification accuracy (Genuine Acceptance Rate (GAR) or True Positive Rate (TPR)[2]) obtained at 1% False Acceptance Rate (FAR)[3]. All experiments are performed using the PyTorch framework on a Nvidia 1080Ti GPU. In order to ensure consistency across analysis and fair comparisons, fixed protocols have been used for all experiments. For training from scratch, each model has been trained for 50 epochs with a batch size of 64 using the Adam optimizer. Model files will be released for reproducibility.

Network Details: Four deep learning based face recognition networks are analyzed to understand the in-group effect due to different attributes:

1. LightCNN-9 [192]: trained from scratch;

2. LightCNN-29 [192]: pre-trained on the CASIA-WebFace [200] and MS-Celeb-1M datasets[4] [62], which contains 10 million images of nearly 10 thousand identities in the training set;

3. ResNet-50 [65]: trained from scratch and pre-trained on the VGG-Face2 [25] and MS-Celeb-1M datasets for face recognition[5]; and

4. SENet-50 [68]: pre-trained on the VGG-Face2 [25] and MS-Celeb-1M datasets for face recognition[6]. VGG-Face2 dataset contains 3.31 million images pertaining to 9,131 identities. Therefore, the two models (ResNet-50 and SENet-50) are trained on over 13 million images, the largest corpus used for training a publicly available face recognition model.

[2]f$GAR = TPR = \frac{TruePositive}{TruePositive+FalseNegative}$
[3]$FAR = \frac{FalsePositive}{FalsePositive+TrueNegative}$
[4]Link to model: https://github.com/AlfredXiangWu/LightCNN. MS-Celeb-1M dataset has now been retracted.
[5]https://github.com/cydonia999/VGGFace2-pytorch
[6]https://github.com/cydonia999/VGGFace2-pytorch

4.4 Datasets and Protocols

The dataset and corresponding protocol details for each case-study are given in the following paragraphs.

Datasets Used for Case-study-1 (Effect of Ethnicity): The behavior of deep learning networks is analyzed with respect to two groups: *Group-A (Whites)* and *Group-B (Blacks)*[7] Three datasets containing colored images (RGB) are used for performing the said analysis:

1. CMU Multi-PIE dataset [159]: Over 44K images of 336 subjects are selected which correspond to frontal face images having illumination and expression variations.

2. Craniofacial Longitudinal Morphological (MORPH) Album-2 dataset [151][8]: Over 52K frontal face images pertaining to 10,409 subjects are used.

3. Racial Faces in the Wild (RFW) [181] is an unconstrained dataset containing images of different ethnicities, collected from the Internet. Its test set contains labeled images *across ethnic groups*. For analysis, the test set containing 10,196 Group-A and 10,145 Group-B face images of 2,959 and 2,995 subjects, respectively, has been used.

The CMU Multi-PIE dataset is used for analyzing the networks for Group-A, while the MORPH dataset is used for Group-B. Both the datasets are captured in constrained settings, and for both, 70% of the total subjects are used to create the training (or fine-tuning) partition, while the remaining 30% subjects form the test set. This ensures disjoint training and testing partitions, in terms of images and subjects. Due to the availability of limited labeled training data in the RFW datset, only the test data is used to evaluate our hypothesis. Therefore, unless explicitly specified, analysis is drawn using the CMU Multi-PIE and MORPH datasets. Figure 4-3 presents sample images from the three datasets.

[7]For this study, we use the existing classification of ethnicity, such a lighter skin and darker skin to understand machine learning based face recognition models. The two classes are referred to as Group-A and Group-B in the manuscript.

[8]We acknowledge that the MORPH dataset is strongly biased towards specific ethnicities. We have only used the dataset for academic research purposes to study algorithmic bias.

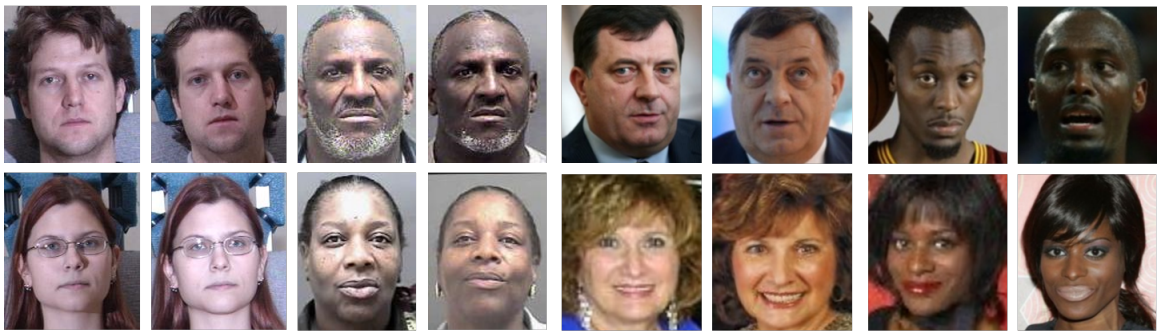


Figure 4-3: Datasets used for understanding ethnicity bias: (i) CMU Multi-PIE (Group-A), (ii) MORPH (Group-B), and (iii) RFW: Group-A and Group-B.

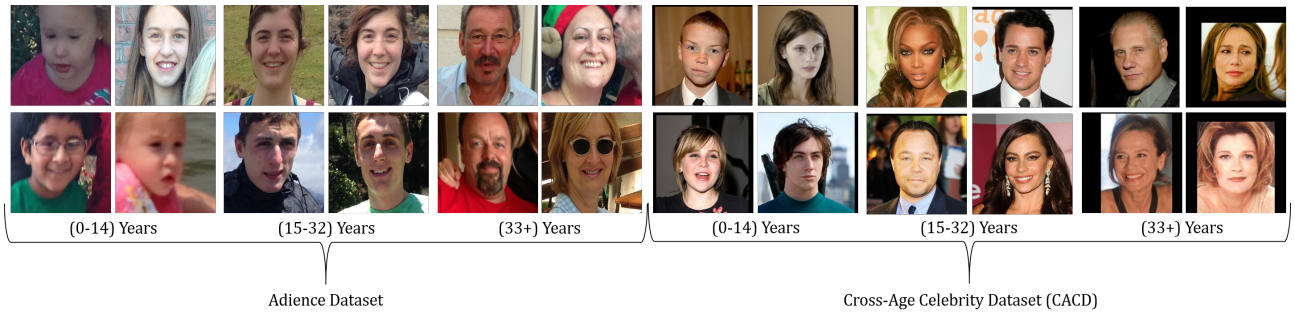


Figure 4-4: Sample images of the two datasets used for analyzing the effect of varying age on deep learning networks.

Datasets Used for Case-study-2 (Effect of Age): Deep learning networks are analyzed for face recognition with respect to three age groups: (i) (0-14) years, (ii) (15-32) years, and (iii) (33+) years. To the best of our knowledge, there does not exist any large scale dataset containing the identity and age information for such a large range. IMDB-Wiki [154] is a large dataset of face images with age and identity information, however, it has been observed that the dataset contains erroneous age information. Therefore, in this study, we have combined face images from two datasets:

1. Adience dataset [49]: The dataset contains over 26,000 face images of 2,284 subjects in the age-range of [0-60+] years.

2. Cross-Age Celebrity dataset (CACD) [30]: The CACD dataset contains over 160,000 face images of individuals belonging to the age range of 14-62 years.

Both the datasets contain images collected from the Internet, having pose, illumination, and back-

ground variations. Figure 4-4 presents sample images from both the datasets, displaying variations in age. In our experiments, the CACD dataset has been used for training (or fine-tuning) the deep learning networks. However, since it does not contain subjects below the age of 14 years, 70% of the Adience dataset has also been used for training. Thus, the training or fine-tuning set is a combination of images from CACD and Adience datasets. The remaining 30% of data pertaining to the Adience dataset is used for testing. Mutual exclusion (for both images and subjects) is ensured between the training and testing partitions.

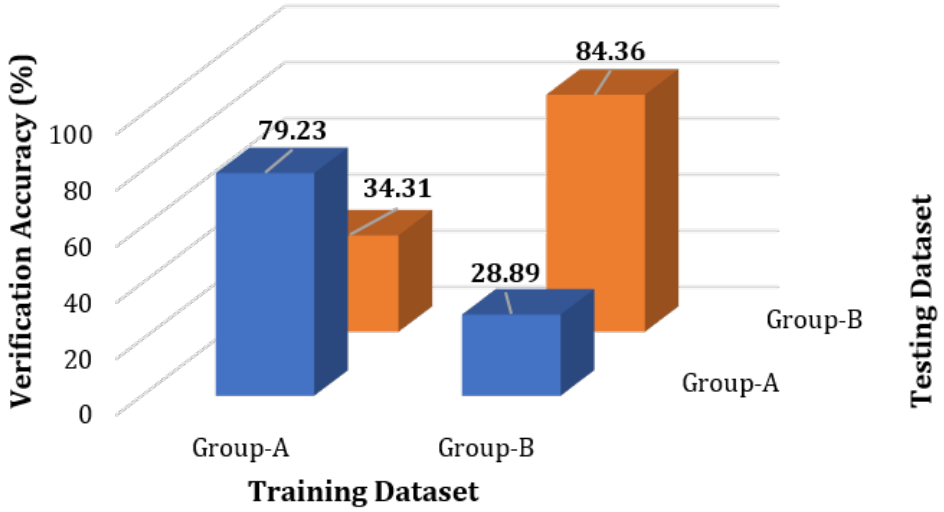
4.5 Does Deep Learning Encode In-Group Bias with respect to Ethnicity?

The *in-group effect* in humans has been well established, particularly for ethnicity, where we are able to recognize people of our own ethnicity more easily as compared to individuals of another ethnicity [55, 113]. Also referred to as the *own-race bias* or *cross-race effect*, it has extensively been studied in humans via cognitive, behavioural, and neuro-imaging studies. Depending on the ethnic group of the individual and the person to be identified, different facial regions have shown more contribution during the decision making process. In this study, we understand the behaviour of deep learning models to understand if they also focus on specific facial regions based on the ethnicity, as observed in humans.

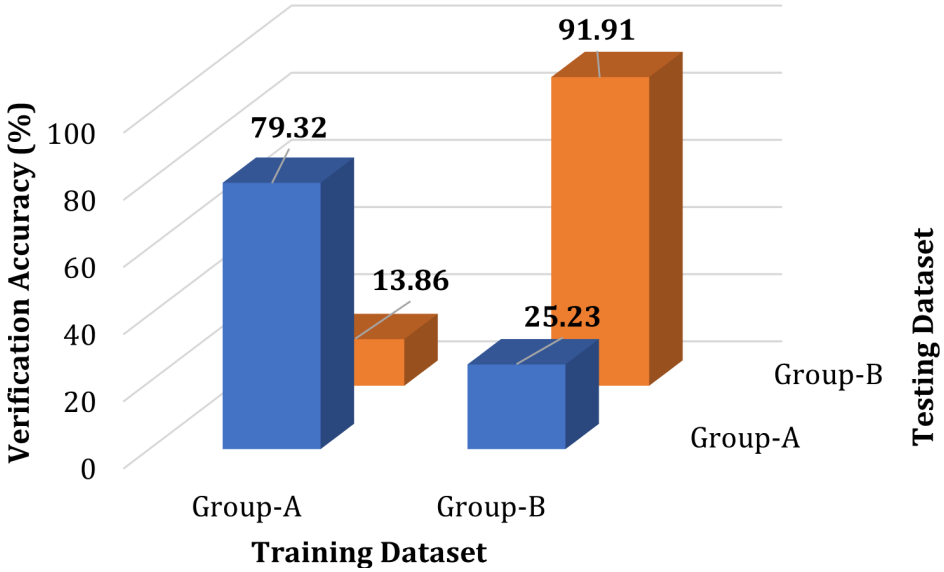
4.5.1 Are Deep Learning Networks Prejudiced?

Experiments 1-10 (Table 4.1) are performed to analyze the presence of prejudice in deep learning networks via class activation maps and feature visualizations. LightCNN-9 and ResNet-50 models are trained from scratch on data belonging to different ethnic groups to understand if faces of different groups are learned in a similar manner, or does the model utilize ethnic group-specific features for identifying individuals.

Figure 4-5(a) presents the GAR values obtained at 1% FAR with LightCNN-9 models, for the two ethnic groups. The model trained on face images of Group-A only reports an accuracy of 79.2% on the testing set of same ethnicity (Group-A). On the other hand, the model trained on



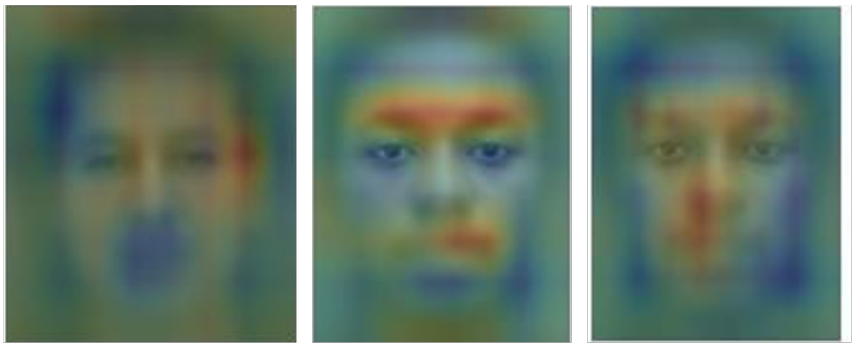
(a) LightCNN-9



(b) ResNet-50

Figure 4-5: Verification accuracy (%) at 1% False Acceptance Rate using (a) LightCNN-9 and ((b) ResNet-50 models. Models trained on images pertaining to a particular group demonstrate poor performance on face images belonging to the other group.

Group-B faces, tested on Group-A individuals, obtains a GAR of 28.9%. Similar trends are observed when testing group-specific models on data belonging to Group-B individuals only. The LightCNN-9 network trained on faces of Group-B subjects only reports a GAR of 84.3% on samples of Group-B (same ethnicity), while the network trained on Group-A faces achieves a GAR of 34.3% (different groups). This suggests the presence of group-specific prejudice being encoded in



(a) Group-A (b) Group-B (c) Combined

Figure 4-6: Visualization of salient regions obtained from expts. 1-5 (Table 4.1), where networks are trained from scratch on (a) only Group-A, (b) only Group-B, and (c) both Group-A and Group-B images. Best viewed in color.

the network due to the variations in the input data. Similar trend is observed from the ResNet-50 model as well (Figure 4-5(b)), where models trained on one sub-group demonstrate poor performance on samples from another sub-group. Additionally, we evaluated the performance wherein equal number of subjects, and equal number of images are selected for both the training and testing sets across the two ethnicities being studied (236 subjects in training and 100 in testing for each ethnicity) using the LightCNN-9 model (Experiment 5-6, Table 4.1). At FAR of 1%, we obtain a GAR value of 85.11% on Ethnicity-A, and 61.64% on Ethnicity-B using the network trained on balanced data. This demonstrates the biased nature of the deep networks, and demands dedicated attention to understand bias, and mitigate this effect to obtain unbiased representations. Further, we also evaluated the LightCNN-9 models trained on specific groups on the RFW test set, and observed a similar trend, thus suggesting that the drop in accuracy is due to the in-group effect and not dataset bias. For the RFW Group-A test set, the network trained on Group-B faces demonstrates a decrease of around 22% of its performance in comparison to the network trained on Group-A. We believe that prejudice occurs due to differences in the learned networks for varying input data.

Feature visualizations allow us to compare models beyond the verification accuracy, and analyze the regions being learned by the network to extract features with the highest discriminative information across groups. Figure 4-6(a-b) presents the salient regions used for feature extraction by the learned networks. These are obtained by interpolating the final convolution layer filter responses and super-imposing on the input image. It can be observed that both models focus on

different regions for feature extraction. Figure 4-6(c) shows the visualizations of salient regions learned by a LightCNN-9 model trained on equal Group-A and Group-B faces. The network learns a union of salient regions obtained for each group independently. While for Group-B, salient regions correspond to the lips and above the eyes, Group-A appear to be distinguished on a more holistic view, with more focus on face boundary. *The differences obtained in the feature maps, along with the accuracy variations strengthen the hypothesis that faces belonging to different groups are encoded differently in a classification model, thereby suggesting the presence of in-group effect.*

4.5.2 Do Deep Networks Mimic Humans?

Cognitive studies in humans have analyzed how Group-A and Group-B individuals perform face recognition [51, 67]. Across different studies, it has been observed that Group-B participants focus more on certain facial features such as mouth, lips, and nose while identifying other individuals of the same ethnicity [51]. Group-A subjects used traits such as the iris color, face shape, hair color and texture for describing and identifying other people of the same group [51, 162]. Such studies suggest a higher level of difference between the regions useful for distinguishing between individuals of different groups. For example, information such as the eye color or hairstyle might not be useful for identifying a Group-B individual, whereas the same information could be of utmost importance for identifying an individual of Group-A.

Intrigued by the findings of human behavior, in this study, deep learning based face recognition models are analyzed to investigate if they follow a similar pattern. In order to understand the behavior of deep learning networks, in terms of the useful regions of interest, class activation maps of the LightCNN-9 models trained on data pertaining to Group-A and Group-B for face recognition are analyzed (Experiment 1-10, Table 4.1). Networks trained on a particular group simulate the functioning of the human brain which has been in contact with individuals of a specific ethnic group only, thereby enabling us to understand in-group effect in deep networks. Figure 4-7 and Figure 4-8 present sample mean class activation maps obtained by the LightCNN-9 and ResNet-50 models. Each map corresponds to the mean activation associated with a particular class. It is interesting to note that the activation maps vary significantly between groups, however, demonstrate a similar

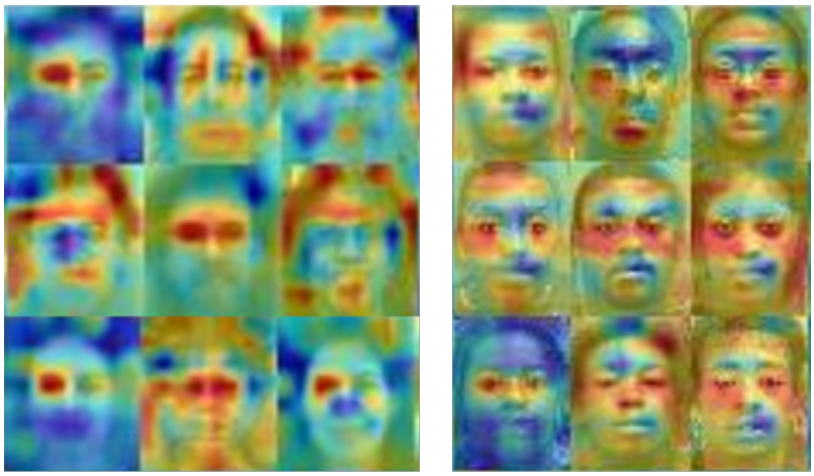
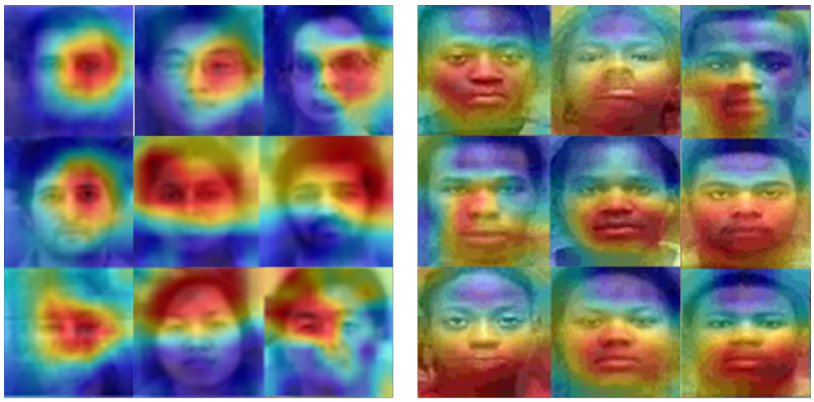


Figure 4-7: Sample average class activation maps from the LightCNN-9 model trained on (a) Group-A and (b) Group-B. Both models appear to focus on different facial regions (Group-A: eyes, Group-B: cheeks/chin). Best viewed in color.

behavior within a particular group. For LightCNN-9, the network trained on subjects belonging to Group-A only, identifies primarily on the basis of the eye region (Figure 4-7(a)). These results are in conjunction with the cognitive studies reported in literature, where researchers have identified the eye color as one of the most identifiable traits used by individuals of Group-A as well [51]. *This suggests a similarity in the behaviour of deep learning networks and the human brain.* Apart from the eyes, we also observe regions around the face being highlighted, which could mean face shape or hair type, both of which have been identified earlier from behavioural studies [51]. Similarly, Figure 4-7(b) presents sample mean class activation maps from the LightCNN-9 network trained on the Group-B dataset. Regions of the lip, nose, and cheekbones appear to contribute more towards face recognition, as compared to other facial regions. Similar results have also been reported in cognitive studies, where Group-B individuals utilized the eyes and lips for describing and recognizing other Group-B subjects [51]. Further, the visualizations can be observed to note asymmetric activations despite the symmetric nature of the human face. As noted in literature as well [195], we believe deep learning models attempt to minimize redundant activations while learning discriminative features. We also obtained similar performance with images flipped across the y-axis. Further, similar patterns are observed with the ResNet-50 architecture as well (Figure 4-8), where models trained on different sub-groups tend to focus on different discriminative re-



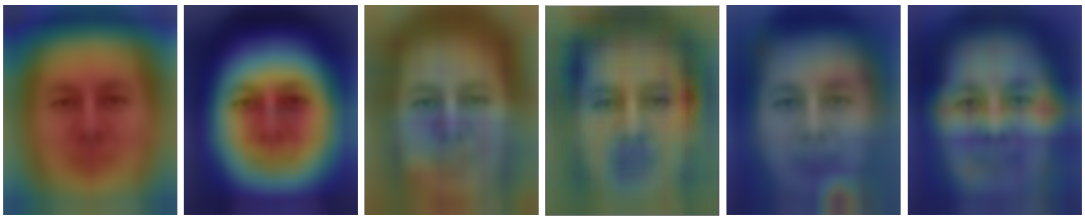
(a) Group-A (b) Group-B

Figure 4-8: Sample average class activation maps from the ResNet-50 model trained on (a) Group-A and (b) Group-B, respectively. Both models appear to focus on different discriminative facial regions (Group-A: eye region, Group-B: chin/lip region). Best viewed in color.

gions. Therefore, *the regions of interest used by models trained on specific groups demonstrate the presence of different discriminative regions, suggesting that networks sub-consciously encode group-specific information, thus resulting in the in-group effect or own-race bias.*

4.5.3 More the Merrier: Does Large-scale Data Help in Mitigating In-group effect?

The presence of in-group effect with respect to ethnicity in individuals is often attributed to limited exposure to other group individuals [94]. In the literature of cognitive studies on humans, researchers have shown reduced in-group effect or the effect of other-race bias upon training subjects to recognize face images of a different ethnic groups. Since deep learning models rely heavily on the training samples, networks trained on large datasets may exhibit a similar behavior. In order to understand whether pre-trained networks suffer from the challenge of bias, deep networks trained on large-scale datasets are analyzed (Table 4.1: Experiments 11-28). The performance of three pre-trained networks - ResNet-50, SENet-50, and LightCNN-29 is studied. Feature maps and class activation maps are also computed to better understand their behaviour. Upon testing on the relatively constrained data belonging to Group-A, ResNet-50 obtains an accuracy of 98.7%, while for Group-B, the network attains a classification accuracy of 96.3%. Similar results are obtained with the SENet-50 network: 98.8% and 96.5%, respectively. As mentioned previously, ResNet-



(i) Pre-trained SENet50 (ii) Pre-trained ResNet50 (iii) Pre-trained LightCNN-29 (iv) Trained LightCNN-9 (v) Fine-tuned SENet50 (vi) Fine-tuned ResNet50

(a) Group-A

(i) Pre-trained SENet50 (ii) Pre-trained ResNet50 (iii) Pre-trained LightCNN-29 (iv) Trained LightCNN-9 (v) Fine-tuned SENet50 (vi) Fine-tuned ResNet50

(b) Group-B

Figure 4-9: Visualization of salient regions obtained from pre-trained, fine-tuned, and trained from scratch networks (Experiment 1-24)). Differences in the *region of focus* are observed across models. Best viewed in color.

50 and SENet-50 are pre-trained on over 13 million face images belonging to the MS-Celeb-1M and VGGFace2 datasets. The VGGFace2 dataset was prepared to explicitly cover a large range of pose, age, and ethnicity in its subjects. In case of LightCNN-29, the model achieves an accuracy of 96.47% and 78.12% on Group-A and Group-B data, respectively. The LightCNN-29 network is pre-trained on the CASIA-WebFace and MS-Celeb-1M datasets. These results suggest that despite pre-training with large-scale data, the lack of variability across ethnicity impacts the performance on the face images of subjects belonging to Group-B.

Similar trend in results is observed on the unconstrained images of the RFW dataset. The pre-trained ResNet-50 network achieves 90.6% on Group-A and 77.3% on Group-B, while the pre-trained SENet-50 model attains an accuracy of 90.4% on Group-A versus 76.4% on Group-B. Similar to the performance on the constrained data, the performance of the LightCNN-29 model reduces to 82.1% and 58.2% for Group-A and Group-B, respectively. The consistent reduced performance on Group-B suggests that the generalization capability of a network depends heavily upon the amount and variability of the training data. Figure 4-9 presents the feature visualizations (region of interest) for the pre-trained networks. Both ResNet-50 and SENet-50 learn feature maps

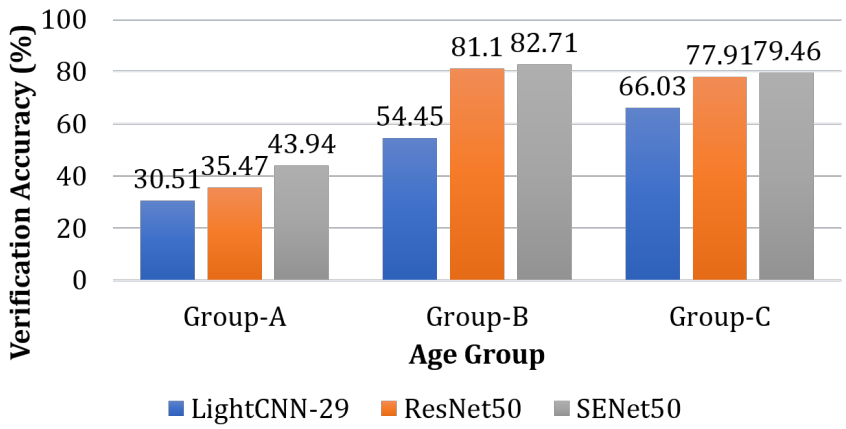


Figure 4-10: Verification accuracy (%) at 1% FAR for three pre-trained networks on the three age based sub-groups.

which cover an entire circular facial region, thus using almost the entire face for recognition, which might result in higher generalization.

Fine-tuning is a common practice for enhancing a network's performance for a chosen task and dataset. Experiments 13-16, 19-22, and 25-28 of Table 4.1 are performed to understand the behaviour of deep networks after fine-tuning. It is interesting to note the transition in *region of interest* from pre-trained networks to fine-tuned networks (Figure 4-9). Initially, for both SENet-50 and ResNet-50, feature maps of both the groups (Group-A and Group-B) are similar, and focus on a circular region covering almost the entire face. After fine-tuning, the updated feature maps appear similar to the ones discussed earlier (Figure 4-6). The feature maps obtained after fine-tuning with images of Group-B are similar to the network trained only on Group-B, demonstrating a shift to only the nose region and face center. In case of Group-A individuals, the networks start focusing on the eye-region only. Beyond visual inference, cosine similarity has also been computed on the feature maps for understanding the similarity quantitatively. The feature maps (as shown in Figure 4-9) are split into eight symmetric patches (similar to Figure 4-13) and the mean pair-wise similarity is computed between two maps. For Group-A, the similarity between the finetuned SENet-50 and pre-trained SENet-50 is 0.8974, while the similarity between the finetuned SENet-50 and the trained LightCNN-9 model is 0.9455. The increase in similarity after fine-tuning suggests a shift in the feature map towards the ones obtained by training with images of Group-A only. Similarly, for Group-B, a variation of 0.9480 (between fine-tuned and pretrained SENet-50) and 0.9706 (between finetuned SENET-50 and trained LightCNN-9) is observed, thus demonstrating a shift in the

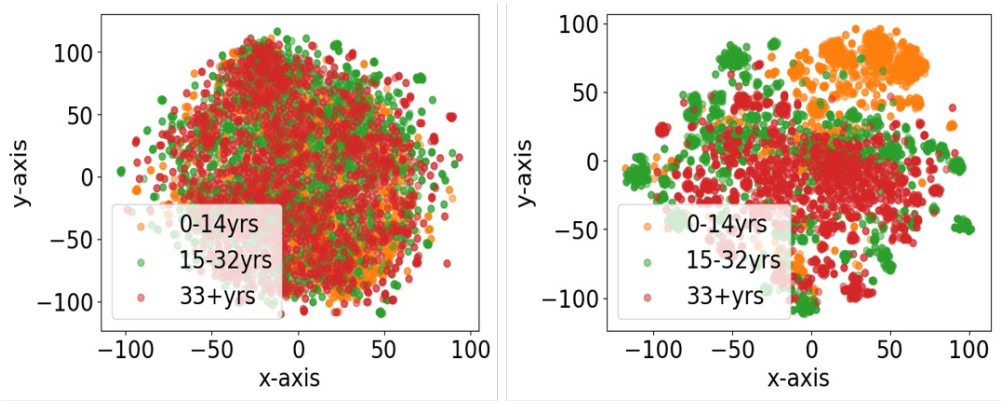
regions of interest towards the ones learned when training with samples of Group-B only. These experiments demonstrate the importance of fine-tuning for specific problems. However, it also establishes that *fine-tuning should only be performed when the test set is highly specific as it might reduce the generalization abilities of the network.*

4.6 Does Deep Learning Encode In-Group Bias with respect to Age?

Similar to the in-group effect due to ethnicity (own-race) bias, researchers have established the presence of in-group effect due to age (*own-age bias*) in human beings [12]. Own-age bias refers to the phenomenon of an individual being able to identify other people of a similar/same age with a greater accuracy (and more easily) as compared to individuals of other ages. Research has focused on understanding the in-group effect with respect to age on three age-groups: children, adults, and older people. While initial studies appeared conflicted in terms of an own-age bias in older people, however, across different studies, researchers have established its presence across different age groups [189]. This is the first research which analyzes and studies bias due to the factor of age in deep learning models. Previously, Kemelmacher-Shlizerman *et. al* [78] presented the MegaFace dataset and studied the impact of age on face recognition with respect to the data used for training a network. However, no analysis of deep models has been performed for the same. This section analyzes the in-group effect with respect to age in deep learning models, specifically in terms of questioning its existence and investigating if distinguishing features vary across age.

4.6.1 How Well Do Deep Networks Recognize Children, Youngsters, and Adults?

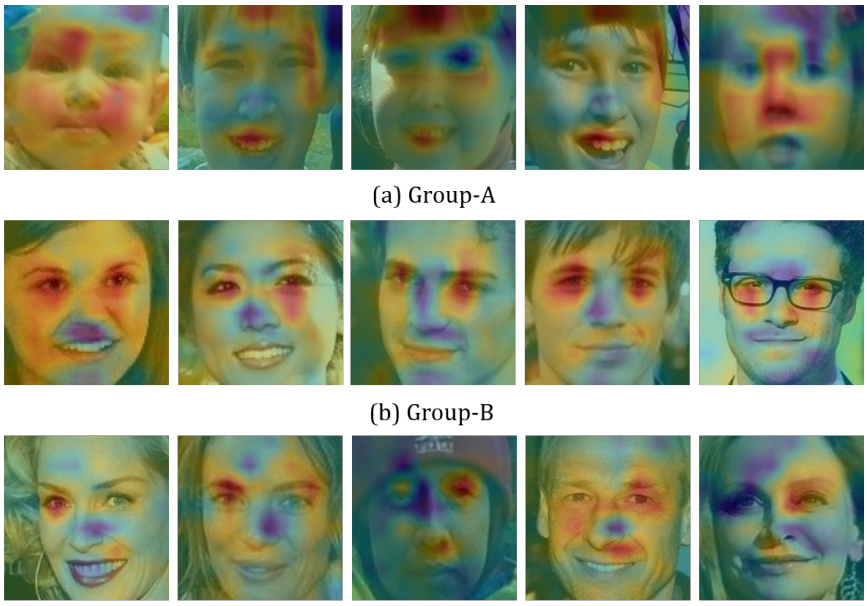
In order to analyze the behavior of deep learning networks in terms of in-group effect with respect to age, a similar set of experiments are performed as the previous case study. Models (both trained from scratch (Table 4.1: experiments 29-31) and pre-trained (Table 4.1: experiments 32-40) are analyzed on face images of varying age-groups. The pre-trained networks of ResNet-50, SENet-50, and LightCNN-29 are evaluated for the three different age-groups: (i) Group-A (0-14 years),



(a) Original Images (b) SENet Features

Figure 4-11: t-SNE visualizations of raw pixels and features extracted from the pre-trained SENet-50 model. 2-D features are obtained using t-SNE and have been plot on the x-axis and y-axis. As compared to the image pixels, the extracted features demonstrate distinction between the children (0-14 years) and other age-groups. Best viewed in color.

(ii) Group-B (15-32 years), and (iii) Group-C (33+ years). The networks which are pre-trained on a large set of data capturing multiple variations, i.e., ResNet-50 and SENet-50 obtained an accuracy of 77.9% and 75.5%, respectively, for the oldest age group of Group-C. Both the networks achieve a slightly higher performance on Group-B, by reporting around 81-83% (Figure 4-10), and less than 45% for the youngest age group (Group-A). On the other hand, LightCNN-29 reports a verification accuracy of 66.03%, 54.45%, and 30.50% for the three age groups, in descending order of age. *The consistent lower face verification performance across different networks for the youngest age-group (Group-C) suggests the presence of an own-age bias in these networks.* Figure 4-11 presents the t-SNE [103] plots for the images (across all three age-groups), and the features extracted from SENet-50. It is interesting to note that features of the youngest age-group (orange colour) appear to form a separate cluster, thereby suggesting that the network is able to distinguish between adults and children, despite being trained for face recognition. This suggests that the network distinguishes between children and adults, however, is not able to distinguish between images of different children. The relatively improved performance of SENet-50 and ResNet-50 models (pre-trained on MS-Celeb-1M and VGGFace2) suggest that similar to the *contact hypothesis* given in cognitive studies, where more exposure to a particular out-group results in increased recognition capabilities, more contact (training) with a sub-group of individuals might result in an improved classification performance for that sub-group. Thus, reinstating the importance of



(a) Group-A

(b) Group-B

(c) Group-C

Figure 4-12: Class activation maps of sample images obtained from the trained LightCNN-9 networks. The models appear to focus on different regions of interest for images belonging to the youngest age group. Best viewed in color.

training data for modeling different covariates of face recognition.

4.6.2 What Does the Model See Based on the Age?

Cognitive research has established the presence of own-age bias, however, limited research has focused on analyzing the specific regions of focus or discriminative facial features across different age groups. While there exist studies which analyze the gaze pattern or scan pattern of individuals belonging to different age-groups, they do not contribute much to the current study, since they do not specifically analyze the discriminative regions of interest. In this study, we analyze the behavior of face recognition models to investigate different regions of interest.

Three LightCNN-9 models are trained, one for each age group, and corresponding class activation maps are presented in Figure 4-12. The class activation maps demonstrate the discriminative regions of focus used by the model for classifying the input images. An interesting trend of activation is observed across the images of different age groups. *Face images of kids (Group-A: first row) appear to have more focus on holistic and soft biometric information such as hair and face shape for classification, along with focus on the lips. On the other hand, models trained on the other age*

groups (Group-B and Group-C) appear to rely heavily on the eye information for classification, along with other facial features. The activations suggest the presence of different discriminative regions across age-groups, thereby suggesting the presence of an in-group effect, especially for young children below 15 years. The presence of different discriminative regions for children provides new insights into the face recognition process, which can potentially open various avenues for novel research directions in both deep learning and cognitive studies.

4.7 Proposed Bias Index

As demonstrated in the previous sections, we aim to understand *where* bias is encoded in deep networks, i.e. what facial regions are focused on by a network for recognition. Based on the observations and insights drawn in this study, we believe that the regions of interest learned by a network can be used to understand and compute the *level of bias* of a given network. As shown in Figure 4-9, models trained on a specific sub-group demonstrate the in-group effect by focusing on group-specific facial regions for recognition (observed via the mean class activation maps). Such models demonstrate highly biased behavior when presented with face images from different sub-groups (e.g. LightCNN-9) and appear to focus on heavily localised facial regions. On the other hand, less biased behavior is observed from models which do not focus on localized regions of interest, and instead utilize almost the entire facial region for feature extraction (e.g. LightCNN-29, SENet-50). The above behavior appears intuitive in nature as well, since a robust facial recognition model should be able to focus on the entire facial region in order to extract meaningful features as opposed to focusing on some localized facial regions only.

Based on the above findings, this research presents a novel bias score metric, termed as the *bias index* to evaluate the *level of bias* of a given facial recognition network by using its mean class activation map. The motivation of using a bias index is to understand if a given model tends to be biased to a particular subgroup or not based on the regions of interest it focuses on to perform classification. The bias index is a combination of two scores: (i) Region of Interest (ROI) score (S_{ROI}) and (ii) Localization score ($S_{Loc.}$), both of which attempt to capture the level of localization for a network's region of focus. The ROI score is a measure of how many facial regions are being focused on by the model, while the localization score is a measure of how scattered the

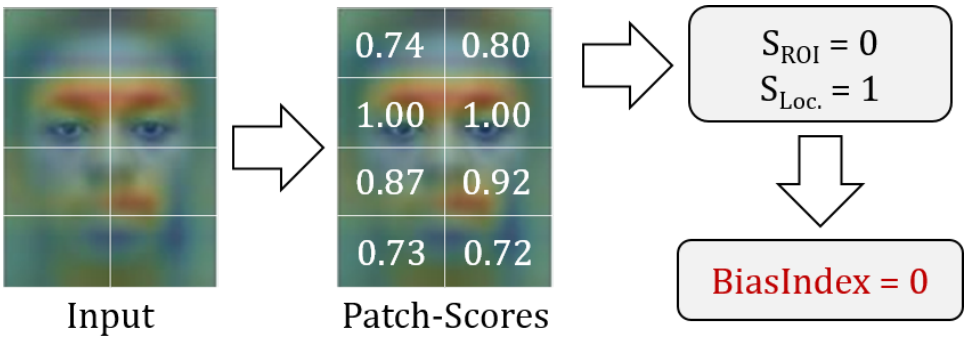


Figure 4-13: The average class activation map obtained for a model is used for predicting its bias index. The image is tessellated into eight patches followed by the computation of patch-scores signifying their importance in the decision making process. The scores are utilized for generating the ROI score and the localization score, which are combined to compute the bias index for the model. Example corresponds to the LightCNN-9 model trained on Group-B (LightCNN-9_B).

facial regions are. The localization score helps in differentiating between models which focus on spatially different regions and models which focus on continual facial regions. For a model $\mathcal{M}$, the bias index is formulated as follows:

$$BiasIndex(\mathcal{M}) = S_{ROI}(\mathcal{M}) + (1 - S_{Loc.}(\mathcal{M})) \tag{4.1}$$

The bias index can take four values (0, 1, 2, or 3) and is used to determine if a network is biased or not in following manner:

$$BiasIndex(\mathcal{M}) = \begin{cases} 3, & Model\ is\ unbiased \\ 1-2, & Model\ is\ moderately\ biased \\ 0, & Model\ is\ highly\ biased \end{cases} \tag{4.2}$$

As mentioned above, the bias index captures the amount of localization in a given network's region of focus. In order to compute the bias index, the mean class activation map obtained by a network is tessellated into equal patches. The patches are divided in such a way that the symmetry across the face is captured and each patch captures a different facial region which has shown to be localized for different groups - forehead, eyes, nose, lips, and the jawline. Thus, the image is divided into eight patches (4x2) such that the resulting patches are able to capture different facial

features without being too coarse (resulting in division of facial features into different patches) or too broad (resulting in multiple features in one patch). The face image is tessellated as shown in Figure 4-13.

A patch-score is generated for each patch which signifies its importance during the decision making process. Mathematically, this is obtained by the ratio of red intensity values in proportion to the green and blue values:

$$Patch - Score_i = \frac{\mu(Red_i)}{\mu(Blue_i, Green_i)} \tag{4.3}$$

where, $Patch - Score_i$ corresponds to the score for the i^{th} patch, μ corresponds to the mean operator, and Red_i/$Blue_i$/$Green_i$ correspond to the values in the red/blue/green channel, respectively for the given patch. The score is computed in such a way since, in a class activation map, red signifies regions of maximum interest, while green and blue donate lesser intensity for contributing in the decision. Therefore in a given patch, if there is a high value of red intensity and lower values of blue and green intensity, the given region is an important region of interest. The patch-scores are then used to compute the (i) ROI score and the (ii) localization score:

- ROI Score (S_{ROI}): The ROI score gives a measure of how many different facial regions are used by the model during decision making. Given the patch-scores obtained above, the number of patches having patch-score greater than 1 is calculated as patch-count. Based on the patch-count, the final ROI score is given as follows:

$$S_{ROI}(\mathcal{M}) = \begin{cases} 2, & if\, patch - count = 8\ (full\ face) \\ 1, & if\, patch - count \geq 4\ (\geq half\ face) \\ 0, & otherwise \end{cases} \tag{4.4}$$

  Localization Score ($S_{Loc.}$): The localization score is used to measure whether the ROIs of the given model are spatially distributed or appear together. Spatially distributed ROIs could result in the model focusing on different facial features, thus resulting in discriminative focus on facial features. On the other hand, if the ROIs appear together, it could result in a holistic view of the face image. Given the Patch-Score for each patch (obtained above), if the scores

Table 4.2: Bias index values for the different trained deep learning models used for studying in-group effect due to ethnicity (Figure 4-9). As before, models are either pre-trained on large-scale datasets (LightCNN-29, ResNet-50, SENet-50) or have been trained from scratch on samples from Group-A or Group-B (LightCNN-$9_{A/B}$, ResNet-$50_{A/B}$, SENet-$50_{A/B}$). Patch-scores, ROI score (S_{ROI}), localization score (S_{Loc}), and the bias index have been tabulated below. The estimated scores resonate the model behavior observed earlier with respect to the face recognition accuracy across different sub-groups.

Model	**Pre-training**	**Patch-Scores**								$\mathbf{S_{ROI}}$	$\mathbf{S_{Loc}}$	**Bias Index**
		P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**			
LightCNN-29	Yes	1.56	1.53	1.12	1.13	0.93	1.01	1.39	1.34	1	0	2
LightCNN-9_A	No	0.91	0.94	0.82	1.07	0.90	0.93	0.81	0.84	0	1	0
LightCNN-9_B	No	0.74	0.80	1.00	1.00	0.87	0.92	0.73	0.72	0	1	0
ResNet-50	Yes	0.71	0.69	1.01	1.11	1.13	1.23	0.66	0.65	1	0	2
ResNet-50_A	No	0.53	0.53	0.73	0.71	0.68	0.70	0.57	0.58	0	1	0
ResNet-50_B	No	0.72	0.70	0.73	0.71	0.82	0.76	0.60	0.60	0	1	0
SENet-50	Yes	0.78	0.78	1.80	1.54	2.01	1.73	0.96	0.91	1	0	2
SENet-50_A	No	0.48	0.51	0.67	0.94	0.68	0.74	0.57	0.67	0	1	0
SENet-50_B	No	0.66	0.66	0.81	0.89	0.91	1.15	0.66	0.70	0	1	0

for three or more neighbouring patches is greater than 1, the $S_{Loc.}$ is set to 0, else it is set to 1.

The above metric has been evaluated on different networks. Table 4.2 presents the patch-scores, ROI scores (S_{ROI}), localisation scores ($S_{Loc.}$), and the bias index ($BiasIndex$) for the different networks used for studying in-group effect due to ethnicity (Figure 4-9). The bias index appears to mimic the behavior observed earlier with respect to the regions of focus, generalizability across different sub-groups, and in-group bias with respect to a specific group. Specifically, a high score is obtained for pre-trained models such as LightCNN-29, SENet-50, and ResNet-50, where lesser variation was observed between the accuracy on different sub-groups, while a low score is obtained on models trained on a specific sub-group. The above behavior further strengthens the utility of the proposed bias index estimation technique for different networks. Such an approach to predict the bias of a network does not rely on the accuracies, or the test set and is rather model specific. The bias index can give an insight into the model learnings. It is designed to reflect the bias-ness of the model irrespective of the dataset. Therefore, the proposed approach is data agnostic. The model could thus be evaluated for bias without computing accuracies and prior to deployment in real-world scenarios.

4.8 Discussions and Summary

Multiple instances reported by researchers have exposed the *hidden biases* of AI systems, creating a need for developing *fairer models*. In this direction, our research analyzes the bias encoded in deep learning networks in the face recognition area with respect to two attributes of age and ethnicity. To the best of our knowledge, this is the first work which analyzes if deep learning networks also exhibit behaviour of *in-group bias*, similar to humans. Additionally, we also attempt to answer *if* and *where* bias is encoded in face recognition models via extensive analysis and visualizations for two attributes specifically: ethnicity and age. Existing studies for analyzing bias have mostly focused on the task of gender classification, and not face recognition. Analysis across multiple deep learning networks suggests the presence of *in-group effect* with respect to ethnic groups (own-race bias) and age groups (own-age bias) in face recognition models. Similar to human behavior, deep learning networks demonstrate a strong tendency of focusing on selected facial regions for a particular demographic, with variations across different groups. Therefore, the models trained on limited variability of data learn models which result in biased behaviour, as they tend to focus on specific regions to perform the task. Upon introduction of data pertaining to a new sub-group, the trained models utilize the same regions to distinguish between samples, which may not be the most effective regions to distinguish the samples and perform recognition, thus resulting in biased behaviour. To the best of our knowledge, none of the existing studies have highlighted that models focus on different facial regions for identifying people of different groups, which could potentially result in biased decisions. Moreover, until now, researchers focused on using large scale datasets in order to capture variability in the data without explicitly ensuring variations across sub-classes such as ethnicity, gender, or age. Recent studies have highlighted the limitation of such an approach since it does not capture the large variation across the different sub-groups both in terms of facial features and accuracy performance. Our analysis also demonstrates the effect of using large-scale data for training. The experimental insights do not recommend it as a complete solution for bias elimination since models trained with large-scale data also demonstrate variation in performance across different sub-groups (as observed in Section V-C). Thus, caution must be taken while curating the training data for deep learning models, in order to incorporate maximum variability possible. Similar behavior is observed for the cross-age study, where

face recognition networks appear to *sub-consciously* encode the age information pertaining to the training samples resulting in varying regions of interest across groups, and presence of *own-age bias*. Lower recognition performance of state-of-the-art face recognition networks on child face images further reaffirms the biased nature of such models with respect to age. Further, based on the observations derived in this research, a novel *bias index* metric has also been presented for estimating the amount of bias in a deep learning based facial recognition model. The performance of the proposed bias index has been demonstrated on the different networks used in this study, where the computed score aligns with the analysis observed via verification accuracy and class activation maps.

Chapter 5

Attribute Aware Filter-Drop for Unbiased Classification

5.1 Introduction

With the advent of machine learning algorithms, automated models have been deployed for various tasks such as object classification for image tagging, resume shortlisting, credit review for loan applicants, facial analysis for targeted advertisements, and spam versus ham classification of e-mails/messages [56, 100, 187]. However, present day Artificial Intelligence (AI) algorithms suffer from the major challenge of biased predictions [7, 76, 80, 119, 135], which often lead to unfair outcomes, thus posing a threat to the widespread use of AI systems. In general, biased systems are usually unfair to a particular section of the society, which can have perilous implications to the structure and balance of the society. For example, gender based biases in automated recruitment drives could result in lesser opportunities to a particular group of people based on a specific attribute [143]. Such instances appear to suggest that *the automated models are not being trained and evaluated on data representative of our society, and rather favor a certain sub-group, resulting in biased behavior.* In computer vision, bias has usually been observed due to skewed training datasets. For instance, for a facial analysis model, the training samples might not be balanced with respect to an attribute such as gender or ethnicity. To this effect, recently researchers have proposed techniques to analyse if existing models are biased [16, 59, 141] or evaluate the fairness of AI models [17, 165], and have also presented unbiased learning strategies [35, 58, 128, 207].

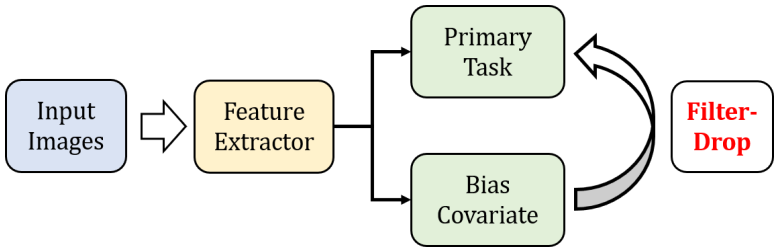


Figure 5-1: Diagrammatic overview of the proposed Filter-Drop technique. A multi-task network is learned to facilitate learning of unbiased features with respect to a given attribute.

In this research, we propose a novel learning algorithm, which at the time of training *unlearns* the dependence of the model on sensitive attributes. These sensitive attributes are referred to as *bias co-variate*. The proposed algorithm, *Filter-Drop*, removes the convolutional filters responsible for encoding a given sensitive attribute (Figure 5-1(b)). For instance, consider the problem of gender prediction from face images with ethnicity as the sensitive attribute. During training, the data contains images belonging to different ethnicities, the proposed algorithms drops filters containing the ethnicity information (sensitive attribute), in order to make the predictions independent of it. The filters to be dropped are learned at the time of training via a multi-task network (Figure 5-1(b)) capable of performing the primary task (gender prediction) along with a secondary task of bias co-variate prediction (ethnicity classification). Filter-drop facilitates the removal of ethnicity features for gender prediction, thus promoting unbiased predictions. The efficacy of the proposed technique is demonstrated on two datasets: UTKFace [211] and FairFace [77] datasets. In order to simulate the real world scenarios, experimental evaluation is performed with skewed data as well as equal data (with respect to the sensitive attribute) to demonstrate the effectiveness of the proposed approach in eliminating the bias present in training samples.

5.1.1 Related Work

The solutions proposed for addressing the challenge of biased predictions can broadly be divided into two categories: (a) learning techniques which design fairer models robust to bias, and (b) bias mitigation algorithms which focus on eliminating the bias from existing models. Detailed review of each category is provided below:

Learning Unbiased Models: In the literature, researchers have explored learning fair models via

adversarial training, learning disentangled representations, or using balanced training data with respect to the protected attribute [202]. Zemel *et al.* [206] proposed a fair representation learning technique such that the input is projected onto another space where the protected attributes are hidden. Das *et al.* [39] proposed a Multi-task CNN for learning a bias-invariant facial analysis model. Zafar *et al.* [205] introduced the concept of fairness of a decision boundary with respect to the protected attribute in terms of disparate impact and disparate treatment. Creager *et al.* [36] proposed a disentangled representation learning technique for obtaining *flexibly fair* features. Kim *et al.* [82] proposed a novel regularization algorithm for learning with biased data by minimizing the mutual information between the feature embeddings and the bias. Nagpal *et al.* [120] proposed a method to drop FCN nodes which learn attribute specific information, for learning unbiased features. Recently, Adeli *et al.* [8] presented a Bias-Resilient neural network, which utilizes adversarial training under two constraints: maximize discrimination for the given task, and minimize the statistical dependence between learned features and the protected attribute. The above models require the protected attribute (e.g. gender/ethnicity) during training, which might limit their usability in real world scenarios.

Bias Mitigation Algorithms: Initial research on bias mitigation focused mostly on the text domain, followed by recent adversarial learning based techniques [104, 184, 204, 207]. Bolukbasi *et al.* [20] proposed a debiasing methodology for eliminating gender bias from word embeddings. Dwork *et al.* [48] proposed learning decoupled classifiers which can be attached to existing black-box models for group-fair classification. Amini *et al.* [11] proposed using a re-weighting based training algorithm for modifying the weights of an existing model. Wang *et al.* [188] proposed an alternative to adversarial learning and inference-time bias reduction, which is domain independent. Recently, Wang *et al.* [180] proposed a reinforcement learning based approach to mitigate bias in face recognition.

Limited algorithms are developed with applicability to the vision domain. This research proposes a novel deep learning algorithm for learning a fairer classification model.

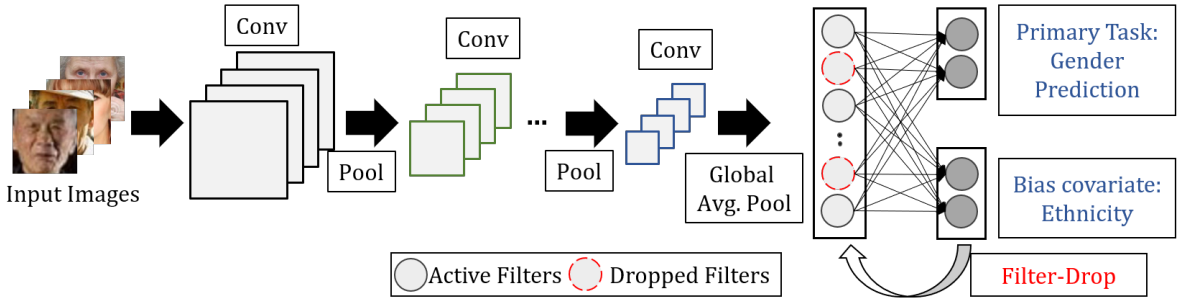


Figure 5-2: Diagrammatic representation of the proposed Filter-Drop algorithm for gender classification, under the sensitive attribute of ethnicity. A multi-task network is learned for gender and ethnicity prediction, such that the filters that are meaningful for ethnicity classification are dropped for gender prediction. The figure has been taken from the published manuscript [120]

5.2 Proposed Attribute Aware Filter-Drop

In this research, we present an approach to perform unbiased and effective classification by learning to drop filters which promote dependency on an underlying attribute. Figure 5-2 presents an overview of the network architecture and the proposed Filter-Drop technique.

5.2.1 Filter-Drop

The proposed concept of filter-drop is similar to that of dropout [167]. Dropout has been used in literature for various applications [132, 194], and several variants of dropout have been proposed in the literature [79]. However, unlike dropout which is performed to improve the generalizability of a network, filter-drop is performed with the specific aim of un-learning. A few filters are dropped or not used for training the subsequent layer, and therefore do not contribute to the final predictions made by the model. The filters to be dropped are learned during training, and the predictions are performed without the dropped filters. In this work, we perform filter-drop before the fully connected (FCN) layer of a network, after applying the global average pooling operator [98]. For an input x, a feature vector is obtained after the final convolution layer ($f(x)$), which is of dimension $d \times m \times m$, i.e., it consists of d filter activations, of dimension $m \times m$. A $d \times 1$ dimension vector is obtained upon applying the global average pooling operator to the feature maps. This is mathematically expressed as:

$$y = \phi(f(x)) \tag{5.1}$$

where, ϕ represents the process of global average pooling. Further, given the output of the global average pooling layer (y), we propose applying filter-drop to obtain y_{drop}, which can mathematically be written as:

$$y_{drop} = m * y \tag{5.2}$$

where m is a $d \times 1$ dimension binary vector, and $*$ refers to element-wise multiplication. If $m = 0$, the filter is dropped and the value is not used for prediction, whereas if $m = 1$, the filter is not dropped. The value of m is determined based on a pre-defined constraint which decides whether a specific feature will contribute towards the final prediction or not. The pre-defined constraint for m can either be defined as random or learned at the time of training to determine the active filters.

5.2.2 Attribute Aware Filter-Drop

As explained in the previous sub-section, the proposed filter-drop can be performed either by using a pre-defined value of m or by defining a constraint to learn m in order to drop the filters intelligently. In this research, *attribute aware filter-drop* is proposed, where the filters containing sensitive attribute-specific information are dropped. This is performed by adding additional constraints during training. In order to learn attribute aware filter-drop, a multi-task network is learned, and predictions are performed for an additional task of attribute prediction, also referred to as *bias co-variate prediction* or the *secondary task*. The loss function ($\mathcal{L}_{Proposed}$) for training a multi-task network via the attribute-aware filter-drop is written as follows:

$$\mathcal{L}_{Proposed} = \mathcal{L}_{Primary} + \mathcal{L}_{Attribute} \tag{5.3}$$

Top 'n' filters which contribute the most to the prediction of the bias co-variate are dropped in order to eliminate the underlying effect of the attribute being predicted. These top 'n' filters are chosen based on the weighted activations for the correct attribute class prediction (a higher weighted activation value refers to more contribution towards the final predicted label). Therefore, the predictions for the primary task are made using the $d - n$ filters, which contain limited

correlation with the sensitive attribute. We can mathematically express this as:

$$m_i = \begin{cases} 0, & \text{if } i \in top\ n(y * W_{True-class}) \\ 1, & \text{otherwise} \end{cases} \tag{5.4}$$

where, m_i corresponds to the i^{th} element of the vector m, $W_{True-class}$ is the final layer weight vector of the true attribute class for the corresponding input x. The above Equation is used to retrieve the top 'n' filters contributing the most towards the true class prediction. Thus, we attempt to eliminate the effect of the sensitive attribute for the primary task by limiting the information used for performing classification.

5.2.3 Unbiased Classification

The proposed attribute aware filter-drop technique is presented in order to learn unbiased representations and thus perform unbiased classification. As shown in Figure 5-2, the model is trained as a multi-task network for the primary classification task and a secondary attribute prediction task. Here, the attribute corresponds to the sensitive attribute corresponding to which a network may be biased. Once the filters having the maximum weights for sensitive attribute prediction are identified, they are dropped to eliminate the effect of the sensitive attribute at the time of the primary classification task. The primary classification task is performed without the dropped filters, both at the time of training and testing. Since the filters are dropped from the penultimate layer, Filter-Drop ensures that the features encoding the sensitive attribute are not used for the primary task. The classification is performed with $d - n$ filters only where n represents the number of dropped filters that have learned the attribute specific information. It is important to note that during the test time, the network is a uni-task network with the objective of performing the primary classification task only (as shown in Figure 5-3).

5.3 Experiments and Implementation Details

The performance of the proposed attribute aware Filter-Drop technique has been evaluated for a facial analysis task, specifically, gender prediction. Given an input face image, a gender predic-

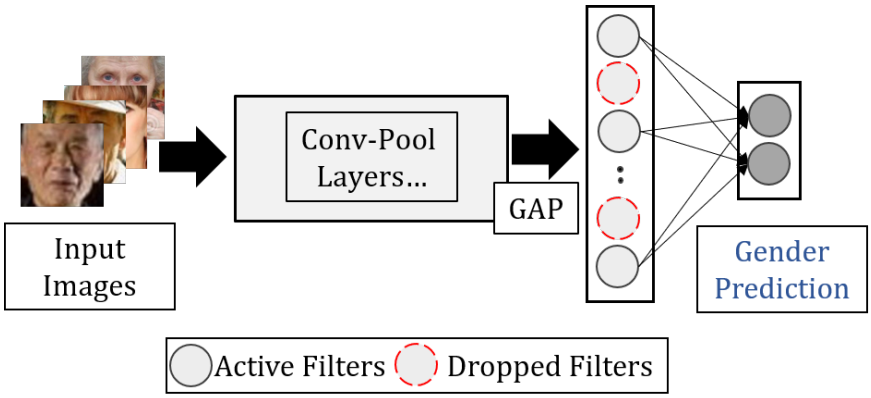


Figure 5-3: Using the attribute aware Filter-Drop model for gender classification during testing. The red filters are dropped and not utilized for the final prediction.

tion model classifies the input either as *male* or *female*. Experiments have been performed on two datasets: (a) UTKFace dataset [211] and (b) FairFace dataset [77]. The UTKFace dataset contains over 20,000 face images from five different ethnicities with large variations across the pose, illumination, resolution, expression etc. Similarly, the FairFace dataset contains face images corresponding to seven ethnicities, collected from the Internet demonstrating a wide range of variations. For experiments, data corresponding to the Asian, Indian, and White ethnicity has been used from both the datasets. Figure 5-4 presents sample face images from the test set of the two datasets.

Two setups have been followed for both the datasets: (i) containing data belonging to the Indian and White ethnicity, and (ii) containing data belonging to the White and Asian ethnicity. In both cases, training has been performed by using equal data from both the ethnicities, and by using training data skewed towards a particular ethnicity. The above two protocols ensure that the proposed attribute aware Filter-Drop technique is evaluated on scenarios resembling the real world having the availability of limited imbalanced training data. For each setup, for the UTKFace dataset, a total of 4000 images have been used for training (containing equal number of male/female samples), while 13,000 images have been used for the FairFace dataset. The test set from the UTKFace dataset contains a total of 2,300 images and the FairFace test set contains 6,000 images, while ensuring equal ethnicity-wise and class-wise distribution for both the datasets.

Ethnicity: Asian

Ethnicity: Indian

Ethnicity: White

(a) UTKFace Dataset [211]

Ethnicity: Asian

Ethnicity: Indian

Ethnicity: White

(b) FairFace Dataset [77]

Figure 5-4: Sample images from the two datasets used for demonstrating variations across different ethnicity groups. The images are captured in unconstrained settings, often showcasing variations along the pose, resolution, lighting, and occlusions.

5.3.1 Implementation Details

Experiments have been performed using the ResNet-50 architecture [65]. As demonstrated in Figure 5-2, the features from the final pooling layer are used for performing two tasks: (i) gender prediction (primary task) and (ii) ethnicity classification (secondary task). A single dense layer is attached to the final feature for the two tasks, respectively. Output of the top 100 filters are dropped during the gender classification for eliminating the component of ethnicity. The model is trained using the proposed attribute aware Filter-Drop technique for 50 epochs, using the Stochastic Gradient Descent optimizer with an initial learning rate of 0.01. After the initial 10 epochs, the dropped filters are estimated, which are updated after each consecutive epoch. The filters obtained at the final epoch are used during test time as well. The network weights are initialized with those learned on the VGG-Face2 dataset [25]. The proposed Filter-Drop has been implemented in the PyTorch framework.

5.4 Results and Analysis

Tables 5.1-5.3 present the results and analysis of the experiments performed using the proposed attribute aware Filter-Drop technique on the UTKFace and FairFace datasets. For all the experiments, the effectiveness of the proposed technique can be observed due to the lower accuracy variation observed between the test set of different ethnicities. Comparison has been performed with the native Softmax loss, termed as 'Traditional', where the ethnicity prediction branch is removed, and the model is trained for the single task of gender classification only.

Analysis Using Equal Training Data: Table 5.1 presents the performance of the algorithms when using equal training data from both the ethnicities. Performance is reported on the test sets of the UTKFace dataset and the FairFace dataset. Since the model has access to equal training data, the performance on the test set is expected to be near similar for different ethnicities. Across different setups, it is observed that the Softmax loss (traditional model) demonstrates relatively higher accuracy variation between the face images of E_A and E_B, while showing a variation of almost 3% between the performance obtained on both the ethnicities. On the other hand, the proposed attribute aware Filter-Drop technique demonstrates lesser disparity between the performance on the two ethnicities.

Table 5.1: Performance of the proposed Filter-Drop technique for gender classification on two datasets when using equal training data from both the ethnicities. Two setups have been followed, where in the first, $\mathbf{E_A}$ refers to the Indian ethnicity, while $\mathbf{E_B}$ refers to the White ethnicity. In setup-2, $\mathbf{E_A}$ refers to the White ethnicity, while $\mathbf{E_B}$ refers to the Asian ethnicity. The proposed Filter-Drop technique demonstrates lower disparity between the performance of different ethnicities.

Dataset	**Algorithm**	$\mathbf{E_A}$	$\mathbf{E_B}$	**Average**
Setup-1				
UTKFace	Traditional	90.69	93.73	92.21
	Proposed	94.52	94.95	94.73
FairFace	Traditional	92.80	94.06	93.43
	Proposed	93.93	94.06	94.0
Setup-2				
UTKFace	Traditional	94.17	93.21	93.69
	Proposed	94.86	94.60	94.73
FairFace	Traditional	94.40	92.00	93.20
	Proposed	94.53	93.06	93.79

Analysis Using Skewed Training Data: Table 5.2 presents the ethnicity-wise performance for two setups on the UTKFace and FairFace datasets for gender prediction. Similar to the previous results, the Filter-Drop technique achieves a lower accuracy variation between the accuracy obtained on the two ethnicities, thus promoting unbiased model learning. In some cases, an accuracy variation of less than 1% (UTKFace dataset) is also observed, thus motivating the usage of the attribute aware Filer-Drop technique for learning attribute (ethnicity) invariant models.

Effect of Number of Filters: Table 5.3 presents the performance of the proposed technique obtained by varying the number of filters to be dropped. The performance is analyzed on the UTKFace dataset having skewed training data for Setup-1 (Table 5.2). The ResNet-50 architecture has 2048 filters in the last convolutional layer, and removing all the filters results in random accuracy for the two class problem (50.00%). Lower variation is observed between the performance on the two ethnicities when removing 100 or 250 filters, as compared to removing 50 or 500 filters. It is our understanding that while dropping 50 filters does not result in the complete elimination of ethnicity information from the model, dropping 500 filters results in the loss of important discriminative information (useful for gender prediction).

Table 5.4 presents the confusion matrix for gender prediction on UTKFace dataset, when

Table 5.2: Gender prediction performance using skewed training data, where only 10% of $\mathbf{E_B}$'s data is used during training. Setup-1 utilizes images from the Indian ($\mathbf{E_A}$) and White ethnicity ($\mathbf{E_B}$), while setup-2 utilizes data from the White ($\mathbf{E_A}$) and Asian ($\mathbf{E_B}$) ethnicity. The proposed technique demonstrates improved performance and less variation across different ethnicities.

Dataset	**Algorithm**	$\mathbf{E_A}$	$\mathbf{E_B}$	**Average**
Setup-1				
UTKFace	Traditional	91.30	93.39	92.34
	Proposed	94.60	94.60	94.60
FairFace	Traditional	91.73	94.40	93.06
	Proposed	93.53	94.26	93.89
Setup-2				
UTKFace	Traditional	94.17	91.91	93.04
	Proposed	94.62	93.82	94.22
FairFace	Traditional	94.66	90.53	92.59
	Proposed	94.66	91.60	93.13

Table 5.3: Gender prediction accuracy (%) on two ethnicities with varying number of dropped filters using skewed training data for Setup-1.

No. of Filters	$\mathbf{E_A}$	$\mathbf{E_B}$
50	94.69	93.91
100	94.60	94.60
250	94.52	94.69
500	94.68	93.73
2048	50.00	50.00

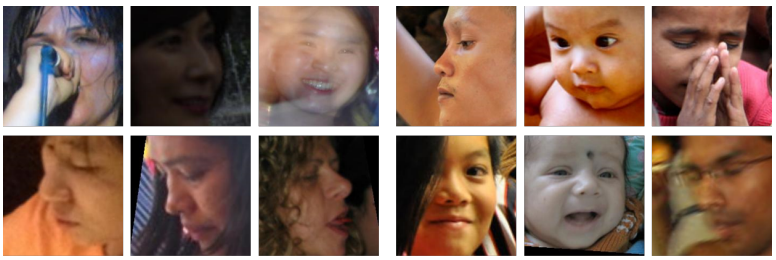
trained with skewed training data on Setup-1. The Filter-Drop technique demonstrates good classification performance across both the classes, where it achieves 95.56% for the female class and 93.65% for the male class. Further, Figure 5-5 presents sample images of the FairFace dataset which were incorrectly classified by the proposed Filter-drop technique. Most of the images suffer from large pose variations, resulting in limited captured face region. Certain images also demonstrate large variations due to the resolution or lighting of the image, thus making the problem further challenging.

Table 5.4: Confusion matrix for gender prediction on the UTKFace dataset using skewed training data with Setup-1 (Indian and White ethnicities). The attribute aware Filter-Drop technique achieves similar performance across the two classes.

		True Label	
		Female	**Male**
Predicted	**Female**	1099	73
	Male	51	1077

5.5 Summary

Deep learning based facial analysis models have been shown to exhibit biased behavior, often resulting in incorrect predictions. Since such models are required to be used in social settings, with access to people from different sub-groups, it is imperative to develop techniques which promote unbiased predictions. To this effect, this research proposes a novel attribute aware Filter-Drop technique for learning features invariant to a given attribute. Filter-Drop utilizes a multi-task network comprising of the primary task and a secondary task for bias co-variate prediction. The primary task refers to the main objective of the network such as facial analysis or object classification, while the secondary task corresponds to the classification of the biasing attribute. For example, for a gender prediction model, the primary task is to predict the gender of the given face image, whereas the secondary task is to predict the biasing covariate, i.e. ethnicity. The proposed technique extracts the top filters containing discriminative information with respect to the secondary task, and focuses on eliminating its features for the primary task. Elimination of the top filters results in the removal of the biasing factor (ethnicity) in the primary task predictions



(a) Female (b) Male

Figure 5-5: Sample images from the FairFace dataset, mis-classified by the proposed Filter-drop technique for gender classification. Large variations due to different covariates of resolution, pose, lighting, and occlusion render the problem further challenging.

(gender). The proposed Filter-drop approach is model and loss agnostic and can be applied to deep learning networks which have a fully connected layer as the penultimate layer. Apart from this, the technique is model agnostic and loss agnostic. The performance of the proposed Filter-Drop technique has been demonstrated on two datasets: (i) UTKFace and (ii) FairFace. Over different experimental setups and varying training data distributions, the proposed technique demonstrates improved performance as compared to the existing algorithm. While current experiments utilize binary primary and secondary tasks, the proposed Filter-Drop technique can also be extended for multi-class problems. This research has been published in the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, and all the figures have been taken from the published manuscript [120].

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Detox Loss for Learning Fairer Facial Analysis Models with Imbalanced Data

6.1 Introduction

Predominantly, bias in machine learning systems is introduced at the data level, due to imbalanced data samples with respect to specific attributes, which is often amplified by the machine learning models [45, 96]. These specific attributes are referred to as *protected attributes* (e.g. gender, ethnicity, or age) and often have a social implication of fairness associated with them [44, 89]. One possible solution to prevent learning biased models which result in unfair predictions is to utilize balanced training data with respect to the protected attributes. Another possible approach, as shown in the previous chapter, is to develop algorithms to unlearn the effect of the protected attribute at the time of making predictions during training itself. Both these kind of methods require an additional label for all the data points i.e. the label of the protected attribute for training puposes. This demand for additional labeled training data is often challenging and cost intensive.

In order to address the data bias and eliminate the need for multi-labeled training data, this research proposes a novel *Detox loss*. As shown in Figure 6-1, despite less number of samples from one sub-group, the Detox loss learns a fairer model, regardless of the training data distribution. The key highlights of this research are:

- For learning fairer models, the proposed Detox loss enforces that learned features must be

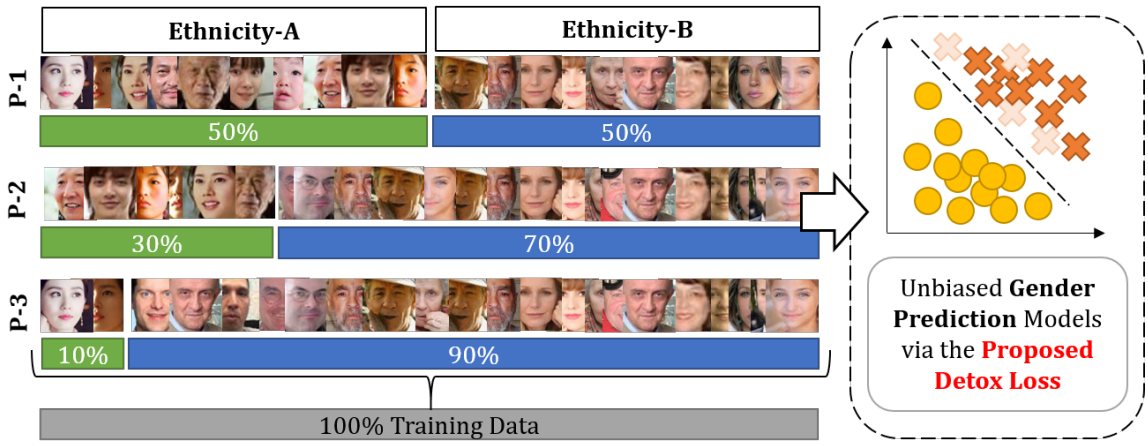


Figure 6-1: The proposed Detox loss learns fairer models even with limited and imbalanced training data (protocols P-1 to P-3) with respect to specific attribute groups. For example, gender prediction models under varying ethnicity groups. Images have been taken from the UTKFace dataset [211].

distinguishable based on the class label only, and not on any additional (protected) attribute. The proposed loss encompasses three additional fairness constraints: (i) the *Exclusion loss*, (ii) the *Inclusion loss*, and (iii) *Feature-distillation loss*, which are used with the traditional cross-entropy loss to achieve the desired effect of fairer classification.

- The Exclusion loss ensures that the learned classifier is discriminative only for the given task, and not over the protected attribute, while the Inclusion loss reduces the spread of the features to ensure data belonging to different protected attributes have similar representations. The feature-distillation loss enables feature learning to be guided by a well-trained pre-trained model by means of knowledge distillation.

- The proposed Detox loss can be used for learning a fairer model from scratch and de-biasing existing pre-trained models as well. Further, the loss *does not require* labeled or balanced data as per the protected attributes to learn fairer classification models.

- The efficacy of the Detox loss has been demonstrated on two facial analysis tasks of (i) age-group classification and (ii) gender prediction. The Detox model has been analysed for learning fairer models as well as de-biasing pre-trained models (LightCNN-29 and ResNet50) with varying the training data distribution with respect to the protected attribute. In the extreme scenario where the training data contains biased data such as 90%:10% across different

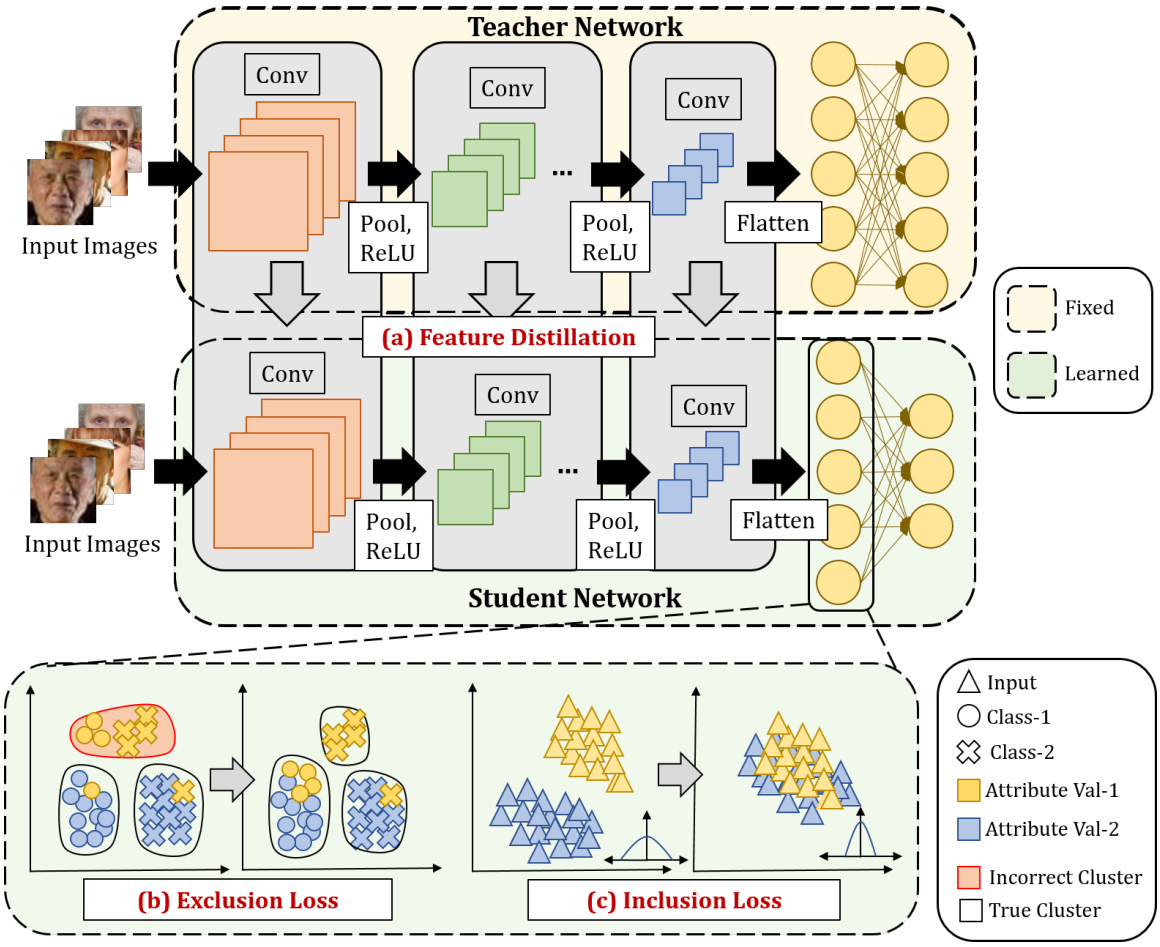


Figure 6-2: Diagrammatic representation of the proposed Detox loss, depicting the proposed (a) Feature Distillation, (b) Exclusion Loss, and (c) Inclusion Loss. The Inclusion loss learns similar features across the protected attribute (e.g. ethnicity), while the Exclusion loss ensures grouping based on class information only. Feature distillation minimizes the distance between the intermediate representations of the teacher network and the student network (trained for a given classification task).

attribute groups, the proposed Detox loss is able to achieve near equal performance on both the groups. Further, comparisons have also been performed with the existing techniques, where the Detox loss achieves state-of-the-art performance.

6.2 Detox Loss: Proposed Unbiased Feature Learning

Figure 6-2 presents an overview of the proposed Detox loss. It consists of three proposed fairness constraints used in conjunction with the traditional Cross-Entropy loss ($\mathcal{L}_{C-E}$): (i) Exclusion loss

($\mathcal{L}_{Exclusion}$), (ii) Inclusion loss ($\mathcal{L}_{Inclusion}$), and (iii) Feature-distillation loss ($\mathcal{L}_{F-Distill}$). Each loss component contributes to the learning of a fairer classification model while applying the following constraints:

- Ensure discrimination based on the class information only and not on the protected attribute ($\mathcal{L}_{Exclusion}$),
- Minimize feature variance across protected attributes for each class ($\mathcal{L}_{Inclusion}$), and
- Learn meaningful features by adapting from a pre-trained classification model ($\mathcal{L}_{F-Distill}$), when available.

Via the combination of the above fairness constraints, the Detox loss facilitates learning of unbiased features irrespective of the sample distribution of the biasing class, and does not require the protected attribute information during training. The proposed loss extends beyond modeling the inter-class and intra-class variations for enhanced classification, and instead utilizes the proposed *Exclusion* and *Inclusion* losses together for learning unbiased features. While the Exclusion loss ensures features are discriminative, the Inclusion loss forces the features belonging to different attributes to be close to one another by minimising the overall spread of features. Since this process happens iteratively, training the model with Detox loss obtains unbiased features. Details regarding each fairness constraint of the proposed Detox loss are as follows:

Proposed Exclusion Loss ($\mathcal{L}_{Exclusion}$): The Exclusion loss aims to learn an *ideal* feature space, where distinction is enforced on the basis of the class information *only*, and not with respect to the protected attribute label. It is our hypothesis that this can be achieved when the model learns 'true' clusters in the feature space, i.e. clusters which contain data belonging to one class only. To this effect, samples which fall into a different class cluster (possibly due to the similar protected attribute) are explicitly pushed towards towards their true class. Thus, the Exclusion loss acts as the first fairness constraint by promoting learning of more discriminative features, based on the class label only. In order to achieve this, the learned features are clustered into m clusters, followed by computing the number of samples of each class in each cluster (Figure 6-2(b)). Here, m must be at least the size of the number of classes. For every cluster, the distance is minimized for all the

samples belonging to the minority class from the mean representation of that class. The Exclusion loss is implemented as follows:

- **Step 1**: Compute clusters $M_1, ..., M_m$ on the features.
- **Step 2**: In each cluster M_i, count the number of samples belonging to each class C, for $C = 1, .., C_n$, where n is the total number of classes.
- **Step 3**: For a cluster M_i, identify the class with minimum samples:

$$C_{min} \ = \ min \ (\text{\# of samples of } C_j); j = 1, .., n \tag{6.1}$$

- **Step 4**: For all samples of class C_{min} in cluster M_i ($\mathbf{X}^{Mi}_{C_{min}}$), minimize the distance between the samples and the corresponding class mean ($\mu_{C_{min}}$):

$$\mathcal{L}_{Exclusion} = \sum_{i=1}^{m} \|\mathbf{X}^{Mi}_{C_{min}} - \mu_{C_{min}}\|_2^2 \tag{6.2}$$

  where, $f(x)$ corresponds to the embedding/representation obtained from the deep learning model for sample x and $\mu_{C_{min}}$ refers to the mean representation of class C_{min}. Mathematically, the class mean can be described as the average of the learned embeddings ($f(x^k_{C_{min}})$) for all samples belonging to the class c_{min}.

Thus, by iterating over all clusters, the Exclusion loss focuses on forming true clusters, i.e. clusters containing samples belonging to one class only. This promotes learning features distinguishable on the class label only, and eliminates protected attribute based distinction.

Proposed Inclusion Loss ($\mathcal{L}_{Inclusion}$): The proposed Inclusion loss acts as the second fairness constraint by modeling the feature space to prevent *sub-conscious* learning of different sub-groups for the given data, by minimizing the overall feature variance. A lower feature variance results in a compact feature space across protected attributes. Coupled with the Exclusion loss, Inclusion loss ensures that features of different sub-groups (e.g. race or age for gender classification) lie in the same feature space, and the learned classifier performs effective and fairer classification despite the protected attributes of the input data. Figure 6-2(c) presents a diagrammatic representation of the

Inclusion loss, where the overall variance of the learned features is minimized. Mathematically, the variance of r points, with mean μ is given as:

$$Var\{\mathbf{X}\} = \frac{1}{r}\sum_{i=1}^{r}(x_i - \mu)^2 \tag{6.3}$$

where, x_i refers to the i^{th} sample of all data points ($\mathbf{X}$), while μ refers to the mean of the data. The Inclusion loss utilizes the above equation as follows:

$$\mathcal{L}_{Inclusion} = \frac{1}{r}\sum_{i=1}^{r}\|f(x_i) - \mu\|^2 \tag{6.4}$$

where, x_i refers to the i^{th} training sample of the dataset ($\mathbf{X}$) containing r samples. $f(x)$ refers to the representation learned by the model and μ refers to the mean representation of the dataset. As described above, $f(x)$ can be the embedding (learned representation) obtained at the penultimate layer of the deep learning based classification model and μ corresponds to the average of the embeddings ($f(x_i)$ obtained on all the samples of the training set. The Inclusion loss also ensures that the group with lesser training samples benefits by being represented in the same feature space as the samples of the other group, by being closer to the overall dataset representation, thus ensuring effective feature learning.

Proposed Feature-Distillation Loss ($\mathcal{L}_{F-Distill}$): As shown in Figure 6-2(a), the proposed Feature-Distillation loss is used as the third fairness constraint to learn from a pre-trained instance of the same model architecture. Traditionally, distillation refers to the process of learning a compact (or smaller) model from a larger original model, capable of replicating the functioning of the original model with minimal errors. Owing to the abundant availability of pre-trained models achieving high performance, the feature-distillation loss facilitates retaining the knowledge of such networks while eliminating their biased behavior (by being trained in conjunction with the other losses). In order to achieve this, the loss between the soft labels of the student network being learned and the teacher network (pre-trained model) is minimized to learn features for accurate classification. The feature-distillation loss is conceptualised by minimizing the distance between the corresponding

feature maps of the teacher and student network. For an input x_i, it is mathematically given as:

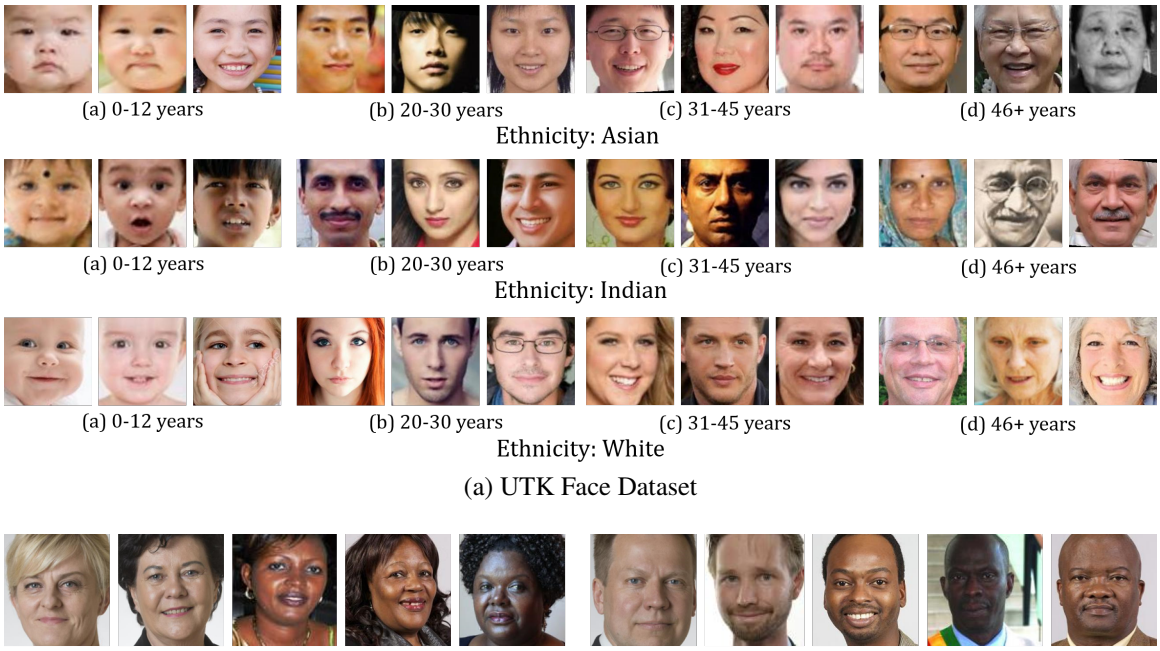
$$\mathcal{L}_{F-Distill} = \sum_{l=1}^{L} \|f_l^T(x_i) - f_l^S(x_i)\|^2 \tag{6.5}$$

where, $f_l^T(.)$ and $f_l^S(.)$ refer to the features extracted by the teacher and student networks at the l^{th} layer, respectively. The loss has two-fold purposes: (i) it enforces meaningful features by learning from a (similar architecture) network pre-trained on a large-scale dataset on a similar domain (e.g. adapting a large-scale face recognition network for a facial analysis task), and (ii) mitigating bias from existing trained models, by distilling onto a new network and optimizing with additional constraints for a fairer model.

Proposed Detox Loss ($\mathcal{L}_{Detox}$): The proposed Detox loss is a function of the three proposed fairness constraints: (i) Exclusion loss, (ii) Inclusion loss, and the (iii) Feature-Distillation loss, used along with the traditional Cross-Entropy loss. The Inclusion loss is complementary to the Exclusion loss since the Exclusion loss enforces distinction on the basis of the class label, while the Inclusion loss minimizes the variance in the training features, thus ensuring data belonging to all sub-groups for a particular class lie in the same space. Thus, the Exclusion loss and Inclusion loss go hand-in-hand and work together to ensure that samples are discriminative on the basis of the class-label only, and features of different protected attributes of each sub-group lie near each other. The Detox loss does not require labeled data with respect to the bias inducing attribute to learn unbiased representations or mitigate the effect of bias. In this research, the detox loss is conceptualised as a weighted sum of the four losses, and given training data $\mathbf{X}$, it is given as:

$$\begin{aligned} \mathcal{L}_{Detox} = \mathcal{L}_{C-E} \; + \; \lambda_1 \big(\sum_{i=1}^{M} \|\mathbf{X}_{C_{min}}^{Mi} - \mu_{C_{min}}\|^2 \big) + \\ \lambda_2 \big(\frac{1}{r} \sum_{i=1}^{r} \|f(x_i) - \mu\|^2 \big) + \lambda_3 \big(\sum_{i=1}^{r} \sum_{l=1}^{L} \|f_l^T(x_i) - f_l^S(x_i)\|^2 \big) \end{aligned} \tag{6.6}$$

where λ_1, λ_2, and λ_3 refer to the regularization constant for each component of the Detox loss. Combined, the proposed loss learns a fairer classification model.



(a) UTK Face Dataset

(b) Pilot Parliaments Benchmark (PPB) Dataset

Figure 6-3: Sample images from the datasets used in this research. (a) The UTKFace dataset demonstrates variations across three ethnicities (Asian, India, White) and different age groups for two genders (male/female). (b) The Pilot Parliaments Benchmark (PPB) dataset contains face images across different skin tones for two gender groups.

6.3 Datasets and Experiments

The Detox loss has been evaluated on two facial analysis tasks: (a) age-group classification and (b) gender classification, under the protected attribute of ethnicity. Analysis has been performed for training models from scratch and de-biasing existing models (LightCNN-29 and ResnNet50) for varying training distributions across the protected attribute (50-50%, 30-70%, 10-90%). Details regarding the datasets, experimental protocols, and the implementation details are provided below.

6.3.1 Datasets and Procotols

Figure 6-3 presents sample face images from the two datasets used: UTKFace dataset [211] and PPB dataset [24]. Details regarding each dataset are as follows:

(i) The UTKFace dataset [211] contains over 20,000 face images collected from the Internet, along with various facial attributes. Each image is annotated with its gender (male/female), eth-

nicity (White/Black/Asian/ Indian/other), and age ([0-116]), with most of the images belonging to White adults. The UTKFace dataset has been used for two facial analysis tasks: (a) age-group classification and (b) gender prediction. Both the tasks have been analyzed under the protected attribute of ethnicity. Similar to Nagpal *et. al* [120], we define a restricted protocol such that each class contains equal training and testing images, and equal images across the protected attribute as well (ethnicity).

- Age-group classification: In this study, four age groups have been considered: (a) 0-12 years, (b) 20-30 years, (c) 31-45 years, and (d) 46+ years. Due to the varying number of samples across different age-groups in the dataset, some age-groups could not be included in this task. Training and testing have been performed on 1600 and 1760 images, respectively.

- Gender prediction: Here, it refers to a two-class problem: male or female. Training and testing have been performed on 4000 and 2300 images, respectively.

We extend the two protocols defined by Nagpal *et. al* [120] for the two tasks mentioned above. Here, regardless of the facial analysis task, images have been taken from two ethnic groups in each protocol. The protocols are:

- **Protocol-1:** Analysis on face images from Indian (E-A) and White (E-B) ethnicity, and

- **Protocol-2:** Analysis on face images from White (E-A) and Asian (E-B) ethnicity.

Balanced accuracy (mean class-wise accuracy) has been reported for the experiments.

(ii) Pilot Parliaments Benchmark (PPB) Dataset [24] is a benchmark dataset for gender prediction from face images. It contains 1270 images of 566 females and 704 males across six skin shades varying from lighter to darker. Consistent with the existing protocol [8], experiments have been performed using five-fold cross validation, where four folds are used for training and the fifth fold is used for testing the model. Balanced accuracy (mean class-wise accuracy) and F-1 scores have been reported for the experiments and comparison has also been drawn with the state-of-art methods.

6.3.2 Experimental Setup

The performance of the Detox loss has been analyzed via two sets of experiments:

- **(a) Learning a Fairer Classification Model:** Experiments have been performed to evaluate the effectiveness of the proposed Detox loss for learning an unbiased classification model for the different tasks discussed above.

- **(b) De-biasing Pre-Trained Classification Models:** Experiments have been performed to de-bias pre-trained classification models for fairer performance.

The UTKFace dataset has been used for the two sets of experiments given above, wherein, data pertaining to two ethnicities (E-A, E-B) has been used for training and testing. In order to simulate the real world scenario of varying training data distributions, experiments are performed with varying proportion of E-A and E-B face images in the training set (specifically, 50%, 30%, and 10% data corresponding to E-B and the remaining to E-A). Evaluation is performed on a fixed set of data containing equal samples from both ethnic groups. On the other hand, the PPB dataset has been used to evaluate the effectiveness of the Detox loss with respect to updating existing classification models and performing comparison with the state-of-the-art techniques.

6.3.3 Implementation Details

For all models, training has been performed with a fixed batch-size of 60 samples, using the Stochastic Gradient Descent optimizer with an initial learning rate of 0.01, which is reduced by a factor of 10 whenever the loss plateaus, for 100 epochs. Grid search has been performed for obtaining the final hyper-parameters (weights for each term in Equation 6.6) as follows: $\lambda_1 = 1e-3$, $\lambda_2 = 1e-3$, and $\lambda_3 = 1e-5$. The Detox loss has been implemented in PyTorch and trained with two Nvidia GTX 1080Ti GPUs.

While training from scratch, the LightCNN-29 [192] architecture is adopted for different facial analysis tasks. Further, the feature-distillation loss is omitted when training a network from scratch. For experiments on pre-trained networks with the UTKFace dataset, a LightCNN-29 model pre-trained on the CASIA Webface dataset [192], and a ResNet-50 model pre-trained on VGGFace2 and MS-Celeb-1M datasets [25] have been used. The LightCNN-29 architecture contains 29 layers

including convolution and max-feature-map (MFM) layers, while the ResNet-50 model contains 50 layers including residual blocks. For experiments on the PPB dataset, as per the protocol [8], VGG-16 [133] and ResNet-50 models have been used (pre-trained on the ImageNet dataset [90]). The Exclusion and Inclusion losses are applied on the penultimate layer of each network, and the feature-distillation loss is applied on every third layer of the model. For the exclusion loss, the K-means algorithm has been used for creating the clusters.

6.4 Results and Analysis

Facial analysis experiments have been performed to evaluate the Detox loss for (i) learning an unbiased classification model and (ii) mitigating bias from existing classification models: LightCNN-29 and ResNet50 (Tables 6.1-6.2). Two facial analysis tasks have been studied: (i) age-group and (ii) gender prediction from face images, under the protected attribute of ethnicity. Experiments have been performed on the UTKFace dataset [211] and the Pilot Parliaments Benchmark (PPB) dataset [24]. For the UTKFace dataset, in each experiment, two ethnic groups (E-A/E-B) have been considered, and different training protocols have been followed having varying training distributions of the two ethnic groups (50-50%, 30-70%, 10-90%). Classification performance obtained on the test set of each ethnicity (E-A/E-B) has been reported. The Detox loss has also been analyzed via an ablation study to understand the contribution of each fairness constraint of the proposed loss (Table 6.3). Comparison has also been drawn with the state-of-the-art algorithms on the standard PPB dataset (Table 6.5) and other recent techniques in literature (Figure 6-4). Detailed discussion and analysis is as follows:

6.4.1 Detox Loss for Training an Unbiased Model

The Detox loss achieves improved performance for age classification, both in terms of reduced accuracy difference between the face images of different ethnicities and average classification accuracy (Table 6.1). For example, when trained with only 10% data of E-B, the base model (Softmax) demonstrates an accuracy variation of almost 5% (protocol-2) across the test sets of E-A and E-B, whereas the Detox loss achieves a difference of around 0.12% for the same setting. Similar behavior can be observed across the two protocols, with two ethnicities, and varying training

data distribution. Similar performance is also obtained for gender prediction, wherein, Table 6.2 presents the gender prediction accuracies of the proposed loss along with the comparative results. The Detox loss is able to learn accurate and unbiased prediction models even with training data containing only 10% of a particular protected attribute. For example, on protocol-1, the Detox loss achieves an accuracy of around 85% across both ethnicities, with standard deviation $< 0.5\%$ across the ethnicities, when trained with only 10% E-B's data. The overall enhanced performance and the lower accuracy variations across the protected attribute (ethnicity: E-A/E-B) promotes the usage of the Detox loss for learning fairer models. Figure 6-4(a) presents the loss values obtained during model training (protocol-1), demonstrating successful convergence.

6.4.2 Detox Loss for Debiasing Pre-trained Models

Table 6.1 can also be analyzed to observe the behavior of different pre-trained models (LightCNN-29 and ResNet-50) with the Detox loss for age-group prediction. For both the models and across both experiments, the Detox loss achieves fairer models with respect to the protected attribute. For example, despite having only 10% training data corresponding to a particular ethnicity (E-B), the Detox loss achieves minor accuracy variation on both the ethnicities (E-A and E-B) across both the experiments (e.g. for protocol-2: Detox (LightCNN): 78.18%, 78.63%; Detox (ResNet): 78.86%, 78.18%). Similar improvement is also observed for the task of gender prediction (Table 6.2). Further, the Detox loss is able to maintain a low accuracy difference ($<$1%) even with extreme training distribution (90%:10% of E-B:E-A). We believe that overall feature variance minimization and cluster modeling enables learning of a fairer classification model, while feature-distillation enables relevant feature retention.

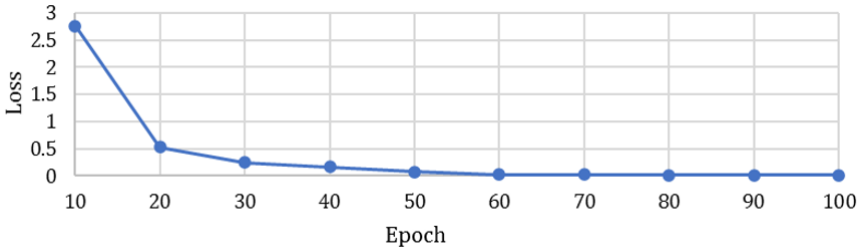
Additionally, we also compare with a recent technique, *Filter-Drop* (Nagpal *et al.* [120]) for gender prediction on both the protocols using the ResNet-50 architecture as backbone for 50%:50% and 90%:10% data distribution. Figure 6-4(b-c) demonstrates that the proposed Detox model achieves comparable performance with Nagpal *et. al* [120]. It is important to note that unlike the proposed Detox loss, the Filter-Drop technique utilizes the protected attribute to train the classifier, and thus require additional labels of the data (class label and biasing variable label) in order to learn a fairer model.

Table 6.1: Balanced accuracy (%) of the Detox loss on age-group classification of face images. Balanced accuracy for each ethnicity (E-A, E-B) have been provided, where the Detox loss achieves less variation across ethnicities as compared to the traditional Softmax loss. Two protocols have been followed: protocol-1 (E-A: Indian, E-B: White) and protocol-2 (E-A: White, E-B: Asian). Experiments are performed by varying the percentage of E-B images in the training set (50%/30%/10%).

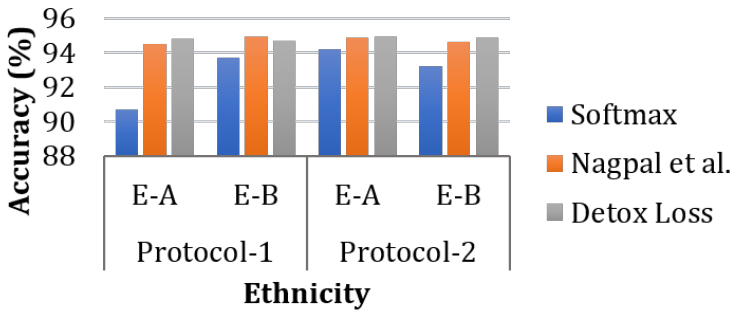
Algorithm	**Protocol-1**						**Protocol-2**					
	50%		30%		10%		50%		30%		10%	
	E-A	E-B	E-A	E-B	E-A	E-B	E-A	E-B	E-A	E-B	E-A	E-B
Learning Unbiased Models												
Softmax	57.04	59.77	58.18	55.68	58.30	55.34	61.25	61.81	59.20	64.54	60.11	55.0
Detox	58.40	59.43	59.43	59.09	60.90	60.90	64.20	65.11	64.20	64.65	61.81	61.93
De-biasing Pre-trained Models												
Model 1: Pre-trained LightCNN-29												
Softmax	68.97	72.84	68.29	71.02	72.38	70.90	77.38	79.09	79.54	75.68	79.77	75.68
Detox	71.02	73.40	72.04	73.86	73.29	73.06	78.40	79.60	78.06	77.61	78.18	78.63
Model 2: Pre-trained ResNet50												
Softmax	68.10	72.50	68.97	72.04	70.45	69.09	75.79	77.38	76.26	72.84	79.43	75.79
Detox	72.15	74.54	72.50	74.20	73.18	73.40	79.30	78.60	78.06	77.95	78.86	78.18

Table 6.2: Balanced accuracy (%) for gender prediction, under the protected attribute of ethnicity. Two protocols have been followed with two ethnicities (E-A, E-B), where each protocol utilizes varying training distribution based on E-B (50%/30%/10%). For protocol-1, E-A: Indian and E-B: White; and for protocol-2, E-A: White and E-B: Asian. Detox loss achieves less accuracy variation across ethnicities for varying training distributions, thus suggesting fairer model learning.

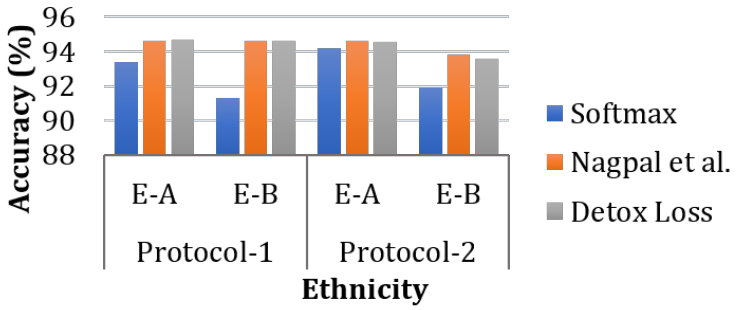
Algorithm	**Protocol-1**						**Protocol-2**					
	50%		30%		10%		50%		30%		10%	
	E-A	E-B	E-A	E-B	E-A	E-B	E-A	E-B	E-A	E-B	E-A	E-B
Learning Unbiased Models												
Softmax	87.21	86.43	88.78	83.91	88.78	84.78	87.30	86.78	85.56	82.26	86.95	82.34
Detox	87.56	87.56	84.17	85.13	84.34	85.82	87.21	88.26	83.47	83.21	83.21	82.43
De-biasing Pre-trained Models												
Model 1: Pre-trained LightCNN-29												
Softmax	90.52	93.73	91.56	94.08	91.30	93.73	94.17	93.30	94.60	92.34	94.60	90.78
Detox	94.52	94.69	94.17	94.34	94.08	94.00	94.43	94.95	95.13	95.73	94.43	94.17
Model 2: Pre-trained ResNet50												
Softmax	90.69	93.73	93.82	91.91	93.39	91.30	94.17	93.21	94.0	92.60	94.17	91.91
Detox	94.78	94.69	94.34	94.08	94.69	94.60	94.95	94.86	94.60	94.60	94.52	93.56



(a) Loss Convergence



(b) 50%:50% Protocol



(c) 90%:10% Protocol

Figure 6-4: (a) Loss convergence graph of the Detox Loss during training on the UTK-Face dataset for gender prediction. (b-c) Comparative performance on the UTKFace dataset for gender prediction. Existing protocol is used [120], where a training distribution of 50%:50% and 90%:10% is followed for E-A:E-B. Detox loss obtains comparative performance as compared to the existing model.

Figure 6-5 presents the t-SNE visualizations of the features at various epochs of training with the proposed Detox loss for age prediction (protocol-1, 90%:10% data between E-A:E-B). Despite being an inherent face recognition model, the features obtained by the pre-trained LightCNN-29 model at the first epoch have discriminative information with respect to the protected attribute (Figure 6-5(e)). After training with the Detox loss, the features demonstrate a more inclusive behavior, wherein data pertaining to the two ethnicities are intertwined (Figure 6-5(h)), while being discriminative in nature for age-group classification (Figure 6-5(d)). The visualizations suggest

Table 6.3: Ablation study on the Detox loss for protocol-2 of gender classification with 90%:10% (E-A:E-B) training data distribution. Balanced accuracy (%) is reported on the entire dataset (Overall), along with the accuracy gap (Gap), which refers to the accuracy difference between the two ethnicities. A smaller difference suggests a fairer model.

Algorithm	**Balanced Acc.**	
	Overall	**Gap**
Softmax	92.69	3.82
Proposed - $\mathcal{L}_{Exclusion}$	93.12	2.09
Proposed - $\mathcal{L}_{Inclusion}$	93.08	3.74
Proposed - $\mathcal{L}_{F-Distill}$	92.91	3.22
Proposed - $\mathcal{L}_{Exclusion}$ - $\mathcal{L}_{F-Distill}$	93.95	1.82
Proposed - $\mathcal{L}_{Inclusion}$ - $\mathcal{L}_{F-Distill}$	93.04	2.78
Proposed - $\mathcal{L}_{Exclusion}$ - $\mathcal{L}_{Inclusion}$	92.65	3.04
Proposed Detox	**94.30**	**0.26**

that the Detox loss minimizes the feature variance across the protected attribute (ethnicity) while ensuring task-specific distinction.

6.4.3 Additional Analysis of the Detox Loss

Ablation Study: Table 6.3 presents the *ablation study* on the proposed Detox loss for gender classification (training data distribution: 90:10% for E-A:E-B, respectively). The Detox loss achieves an overall classification performance of 94.30% (Table 6.3), which demonstrates a reduction upon the removal of any fairness constraint. Experiments have been performed by removing different components (Equation 6.6) from the Detox loss to evaluate their effectiveness in the final algorithm. Individually, the Detox loss obtains 94.43% and 94.17% on E-A and E-B, respectively. Further, the Detox loss learns an unbiased model (accuracy variation of 0.26%), whereas removing any loss component presents a difference between 2.09-3.82%. The maximum increase in accuracy difference is observed by removing the variance modeling component ($\mathcal{L}_{Inclusion}$) (3.74%). Similarly, removing the two components: $\mathcal{L}_{Exclusion}$ and $\mathcal{L}_{Inclusion}$ presents a maximum drop of 3.04%. Since variance modeling forces the model to learn similar features irrespective of the protected attribute, we believe it has the highest contribution in learning unbiased features.

Impact of Different Classification Losses: The Detox loss has also been evaluated in conjunc-

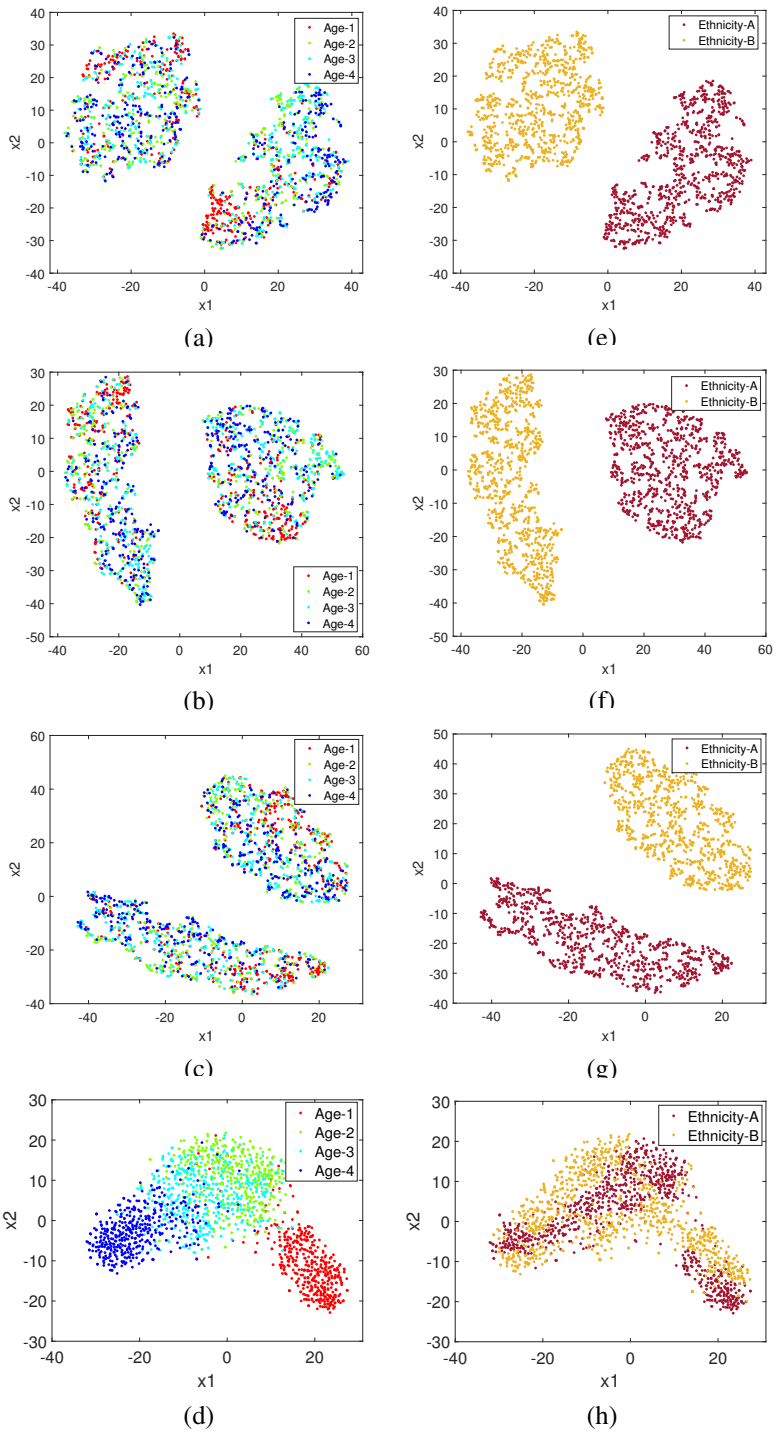


Figure 6-5: t-SNE visualizations of the features obtained while training a pre-trained LightCNN-29 model with the Detox loss for age-group classification. (a-d) present feature visualizations with age-group labels, while (e-h) contain the ethnicity information, at Epochs 1, 10, 20, and 40. The Detox loss learns a discriminative feature space (based on the class label only), without distinguishing between different ethnicity groups. Best viewed in color.

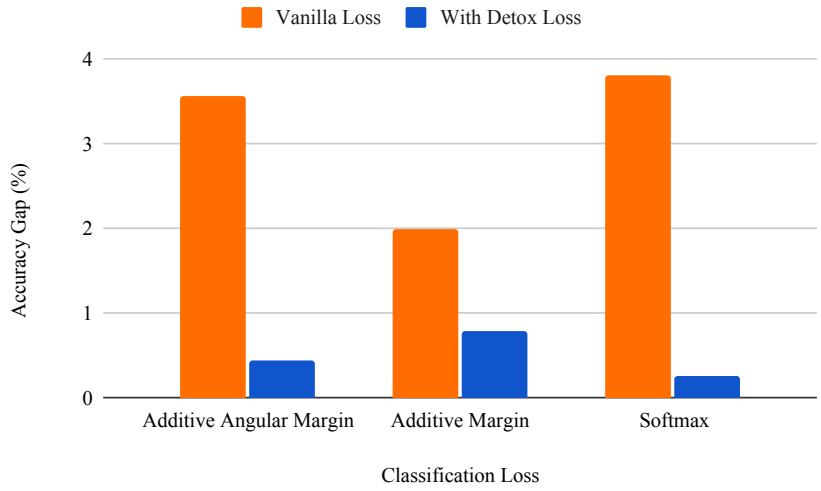


Figure 6-6: Accuracy Gap with Different Classification Losses: Effect of incorporating the detox loss with different classification losses in terms of the accuracy gap (%). A lower gap suggests a fairer model.

tion with two different classification losses, specifically, the Additive Angular Margin loss (ArcFace) [42] and the Additive Margin loss [176]. Similar to the previous analysis, experiments were performed for gender classification (protocol-2, training distribution: 90:10% for E-A:E-B, respectively). Similar to the Softmax loss, for both the classification losses, addition of the fairness constraints via the Detox loss results in a drop in the accuracy gap (resulting in a fairer model) along with an increase in the classification accuracy. Specifically, with inclusion of the Detox loss, the model trained with the Additive Margin loss demonstrates an overall accuracy improvement from 94.21% to 94.39% while presenting a drop in accuracy gap from 2.00% to 0.78%. Similar performance is seen with the Additive Angular Margin loss as well, where the inclusion of the Detox loss results in an overall accuracy improvement from 93.86% to 94.20% along with a drop in the accuracy gap from 3.57% to 0.43%. Fig. 6-6 presents the visual comparison of the accuracy gap (%) with and without the Detox loss for different classification losses. The above analysis thus supports the utility of the Detox loss with different classification losses for learning a fairer model.

Impact of Hyper-parameters: Experiments have been performed to understand the impact of the weight hyper-parameters (Equation 6.6): λ_1 (Exclusion loss), λ_2 (Inclusion loss), and λ_3 (Distillation loss). With the selected hyper-parameters ($[e-3, e-3, e-5]$), an overall accuracy of

Table 6.4: Analysis for number of clusters: Evaluation is performed on the Detox loss for protocol-2 of gender classification with 90%:10% (E-A:E-B) training data distribution. Balanced accuracy (%) is reported on the entire dataset (Overall), along with the accuracy gap (Gap), which refers to the accuracy difference between the two ethnicities. A smaller difference suggests a fairer model.

Number of Clusters	**Balanced Acc.**	
	Overall	**Gap**
2	**94.3**	**0.26**
3	94.6	0.69
4	94.39	0.44
5	93.41	0.3
6	93.82	1.21

94.3% is obtained with a gap of 0.26%. Upon reducing the lambda values to $[e-5, e-5, e-5]$, we observe a slight drop in performance (94.17%) and an increase in the gap (0.7%). Further, on varying the lambda values to $[e-8, e-5, e-5]$, we obtain an overall accuracy of 93.34% and gap of 3.04%. The increase in gap can be attributed to the reduced contribution of the exclusion loss, which promotes distinction on class-label only. Changing the values to $[e-5, e-8, e-5]$ renders an overall accuracy of 93.52% and a gap of 2.18%. This can be attributed to the reduced contribution of the inclusion loss which promotes fair feature learning. Finally, we also evaluated the model with $[e-5, e-5, e-8]$ and obtain an accuracy of 94.12% with a gap of 1.31%.We observe a lesser drop in the gap as compared to other models, which showcases the effect of the fairness constraints for learning an unbiased model.

Impact of number of clusters: As mentioned previously, the number of clusters must be at least the number of classes for the main task. The idea of using clusters is to reinforce that the learned features are distinguishable from one another on the basis of the class labels only rather than the protected attribute. Having clusters larger than the number of classes will result in the creation of multiple pure clusters of the same class, which will be beneficial for learning a fair model since distinction is not being performed based on the protected attribute. We perform experimental evaluation to demonstrate the same, by varying the cluster number from 2 to 6, and observe that we get similar performance across different number of clusters as shown in Table 6.4.

Table 6.5: Comparison of the proposed Detox loss on the PPB dataset for gender prediction. Existing protocol has been followed, where results are demonstrated on two architectures: VGG-16 and ResNet-50. The proposed loss demonstrates improved performance in terms of the balanced accuracy (Bal Acc.) and the F_1 score.

Algorithm	**VGG-16**		**ResNet-50**	
	Bal Acc. (%)	F_1 Score (%)	Bal Acc. (%)	F_1 Score (%)
Vanilla (2021) [8]	94.1±0.2	93.5±0.3	90.7±0.7	89.8±0.7
Zafar *et al.* (2017) [205]	94.3±0.4	93.7±0.5	94.2±0.4	93.6±0.4
Kim *et al.* (2019) [82]	95.8±0.5	95.7±0.5	91.4±0.9	91.0±0.9
Multi-Task (2021) [8]	94.0±0.3	93.4±0.3	94.0±0.3	93.4±0.3
BR-Net (2021) [8]	96.3±0.6	96.0±0.7	94.1±0.2	93.6±0.2
Proposed Detox	**97.4±0.3**	**97.6±0.2**	**96.1±0.4**	**96.6±0.5**

6.4.4 Results on the Benchmark PPB Dataset

Table 6.5 presents the performance of the Detox loss on the Pilot Parliaments Benchmark (PPB) dataset, along with other comparative techniques (owing to the same protocol, results have directly been taken from Adeli *et al.* [8]). The Detox loss demonstrates improved performance across both the architectures as compared to fair representation learning (Kim *et al.* [82]) and bias-resilient neural networks (Adeli *et al.* [8]), both of which attempt to minimize the dependence between the features and the biasing factor. Comparison has also been performed with Zafar *et al.* [205] which focuses on learning fair classifiers. We believe that since the Detox loss focuses explicitly on discriminative feature learning based on the class-label only, it achieves enhanced performance.

Additionally, we also computed the **distance correlation** and **Equality of Opportunity (EO%)**. Distance Correlation measures the statistical dependency between the protected attribute and the learned features, and EO% measures the average gap in true positive rates with respect to different values of the protected variable [8]. For both the metrics, a lower value represents lower dependence between the learned feature and protected attribute. Thus, obtaining a score value of zero is ideal as it implies that the learned feature is independent of the biased attribute. On the ResNet50 backbone, the Detox loss obtains a distance correlation of 0.22, and lies in the top-2 scores, after Kim *et al.* [82], which reports a score of 0.18. However, the balanced accuracy and F-1 score of the Detox loss is at least 4.5% more than Kim *et al.* [82]. The existing state-of-the-art approach [205] in terms of accuracy obtains a score of 0.29, representing higher dependence between the features and protected attribute. Moreover, the proposed model obtains EO% of 1.32%, which is

the lowest in comparison to all the other approaches. The second best EO% is 2.02% achieved by BR-Net [8]. Thus, the lower distance correlation value and EO% further supports our hypothesis of Detox loss learning independent and unbiased features.

6.5 Summary

Developing unbiased training strategies for deep learning based classification models is the need of the hour. This research proposes a novel *Detox loss* for learning fairer classification models. Detox loss learns an unbiased classifier with respect to the protected attributes (e.g. ethnicity or gender), such that the trained model achieves similar performance on different sub-groups. The Detox loss utilizes three novel fairness constraints: (i) Exclusion loss, (ii) Inclusion loss, and (iii) Feature-distillation loss, which are applied in conjunction with the traditional cross-entropy loss. One of the key highlights of the Detox loss is its non-requirement of the protected attributes during training, thus demonstrating wider applicability. Experimental evaluation has been performed on the facial analysis tasks of (i) age-group and (ii) gender prediction. Experiments across varying distributions of training data demonstrate that the Detox loss is able to achieve fair classification accuracies across the protected attribute of ethnicity. Comparison with existing approaches also demonstrate the efficacy of the proposed loss. For example, on the PPB dataset, the proposed loss demonstrates an improvement of around 2% for the balanced accuracy metric and around 3% for the F_1 score metric as compared to the state-of-the-art model. Since the loss is model agnostic, we believe that it can be extended to different classification tasks and thus can have widespread impact and utility.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 7

Conclusion and Future Research

With the advent and evolution of AI technology, face recognition is omnipresent. However, given the fast paced adaptation, some key aspects specifically in order to develop trustworthy facial analysis systems are still less explored or lack dedicated research. This dissertation focuses on two key data centric challenges: (i) face recognition for scarce data such as sketch to photo matching, and (ii) face recognition for biased data. This dissertation presents algorithms for these challenges and the efficacy of the proposed techniques has been demonstrated on several real world benchmark datasets. The key contributions are as follows:

- **Face recognition with scarce data:** The initial contributions of this dissertation include novel transform learning based approaches for face recognition for scarce data, with applicability to sketch face recognition. These approaches are input agnostic and can be applied with different existing feature extractors. The first approach presented is termed as *Deep-Transformer*, wherein data from different domains is transformed into the feature space using the learned transformation and a mapping is learned between the transformed features. The mapping can be either bidirectional or unidirectional depending on the use case. Further, given the challenge of scarce data, with the aim to reduce learnable parameters, and add supervision in the learning process, *Discriminative Shared Transform Learning (DSTL)* is proposed. DSTL learns a shared transform for data belonging to the two domains, while modeling the class variations, resulting in discriminative feature learning. Two models with varying supervision constraints have been presented under the proposed DSTL algorithm:

(i) Contractive Model (C-Model) and (ii) Divergent Model (D-Model). The proposed algorithms have been evaluated on seven datasets for three case studies of sketch to digital image matching, and the proposed models obtain state of art performance thus showcasing the superior performance over existing approaches.

- **Analysis of bias in deep learning models for face recognition and novel bias index:** Understanding state-of-the-art deep learning models is essential to address the challenge of bias in AI, and develop fairer algorithms. Extensive evaluation with respect to ethnicity and age have been performed on a combination of pre-trained and trained from scratch models to understand *if* and *where* bias is encoded in deep learning architectures trained for face recognition. Experimental evaluation showcases that deep learning models focus on different facial regions for different ethnicity or age. A novel *bias index* has also been presented for evaluating a trained model's *level of bias* by analyzing the learned regions of interest from class activation maps.

- **Bias mitigation and unbiased learning strategies for facial analysis:** Deep learning algorithms have shown to exhibit sub par performance on sub-groups despite the high accuracy being reported thus resulting in biased behaviour or unfair performance. In order to address this challenge, we present two different approaches namely: *Filter Drop* and *Detox Loss*. Filter Drop aims to unlearn filters which lead to learning information representing the biasing co-variate. This approach is highly effective but requires the class labels as well as the additional co-variate labels at time of training. Building on the aforementioned loss, the Detox loss is presented which learns the model with three additional fairness constraints such that the learned features are distinguished based on the task label only, while eliminating any distinction based on the biasing-attribute. The efficacy of the proposed algorithms are evaluated on benchmark datasets. Additionally, evaluation is also performed on learning with imbalanced data (90-10%), and similar performance is observed across sub-groups irrespective of the training data distribution.

While this dissertation has made some significant strides in developing trustworthy systems with scarce and biased data, and has advanced the literature with proposed methods leading to improved performance on benchmark datasets, there still remains a lot of scope for future work

and improvement. Following are some identified potential research directions:

- **Generalized sketch based object image retrieval:** This dissertation explored face recognition with scarce data while having applicability to sketch images. With the increase in use of handheld devices and rising human centred computing and design, a lot of hand drawn sketches are being created. Developing algorithms to perform matching of object images belonging to different domains is going to be indispensable. As part of this research, some preliminary experiments and evaluation have been done, however given the wide applicability of such use cases in online search for e-commerce and retrieving similar images, dedicated research efforts will be pertinent.

- **Developing ML/AI techniques for smooth adoption to scarce data for end-to-end pipeline generation:** This dissertation highlighted the importance of heterogenous data from different domains (e.g. sketch) for applications in law enforcement, identity linking, social media tagging, search, etc. Other emerging image domains include data captured in the near infrared (NIR) spectrum or the thermal spectrum. While the wide usage of CCTV cameras has resulted in data captured in the NIR domain (in night-time), thermal imaging cameras are also widely used across industries, especially during the pandemic, to provide a seamless experience to individuals while also checking their body temperature. In future, a single camera system can be used for recognition as well as checking parameters of the individual. Thus, it is important to develop end to end pipelines for these heterogeneous data wherein limited data is available. End-to-end pipeline encapsulates pre-processing, segmentation, detection, recognition, as well as supplementary tasks such as attribute prediction. While research is well explored for various applications such as segmentation, detection, and attribute prediction, etc on RGB data, it is worthwhile to put in research efforts to develop ML/AI techniques which learn from these existing algorithms and can transfer knowledge to build models for scarce data.

- **Data augmentation approaches for scarce and heterogenous data:** Given that the data available for model development and training is scarce, a traditional approach would be to increase the data by applying standard data augmentation techniques such as changing the brightness or contrast. However, images belonging to domains like sketches/ thermal/ near

infrared spectra will not benefit from meaningful augmentation by using such techniques. Thus, it is worthwhile to explore and develop augmentation strategies applicable for varying data types.

- **Handling data bias in scarce data for building trustworthy systems:** Examples of bias have been widely observed in standard machine learning models being used in the real world. As part of this dissertation, we analyse and develop mitigation strategies for bias observed in standard deep learning models. However, on developing models for scarce data, it will be important and interesting to study bias and ensure unbiased learning in these scenarios as well.

- **Bias in object classification :** There has been limited attention given to bias in object data. Researchers have studied and created a preliminary dataset of objects and how they differ across geographies. We have explored [121] preliminary de-biasing techniques to handle such varied data, however the task is challenging due to the number of variations that can exist due to geography as well as socioeconomic status. Developing labeled datasets would be a step forward in this direction to understand the challenges encompassed and if a truly generalizable approach can be developed.

- **Large-scale universal benchmarks for bias estimation to build trustworthy ML models:** Research efforts have been made to understand and mitigate bias as well as learn unbiased models. Researchers have also presented several different metrics to measure bias and evaluate models. Given the widespread applicability and overnight demand for ML models, and the need to learn and develop unbiased models, it is imperative to develop standard benchmarks to report a model's level of bias with standard benchmark metrics. Today, there is no standard way to report bias and measure bias for models. To this effect, such a universal benchmarking system could set standards for performance reporting and result in an eminent metric being reported in future research papers. Such a benchmark would allow fair evaluation of different models along with providing confidence to researchers with respect to the deployability of such models in the real-world setups.

Bibliography

[1] evofit. http://www.evofit.co.uk/. 19

[2] Faces. http://www.facesid.com/products.html. 19, 30

[3] Identi-kit. http://identikit.net/. 19

[4] Verilook. http://www.neurotechnology.com/verilook.html. 33, 35, 56, 60

[5] Propublica study of algorithms. https://tinyurl.com/gvtccpq, 2016. 69

[6] Facial recognition, and bias. https://tinyurl.com/y7rat8vb, 2018. 69

[7] A. Abid, M. Farooqi, and J. Zou. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463, 2021. 97

[8] E. Adeli, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, L. Fei-Fei, J. C. Niebles, and K. M. Pohl. Representation learning with statistical independence to mitigate bias. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2513–2523, 2021. 99, 119, 121, 130, 131

[9] A. Akbari, M. Awais, Z. Feng, A. Farooq, and J. Kittler. A flatter loss for bias mitigation in cross-dataset facial age estimation. In *International Conference on Pattern Recognition*, pages 10629–10635. IEEE, 2021. 74

[10] M. Alvi, A. Zisserman, and C. Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *European Conference of Computer Vision Workshops*, 2018. 14, 74

[11] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019. 14, 74, 99

[12] J. S. Anastasi and M. G. Rhodes. An own-age bias in face recognition for children and older adults. *Psychonomic bulletin & review*, 12(6):1043–1047, 2005. 88

[13] W. Aulsebrook, M. İscan, J. Slabbert, and P. Becker. Superimposition and reconstruction in forensic facial identification: a survey. *Forensic Science International*, 75(2):101–120, 1995. 65

[14] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104:671, 2016. 10

[15] S. Barra, A. Castiglione, F. Narducci, M. De Marsico, and M. Nappi. Biometric data on the edge for secure, smart and user tailored access to cloud services. *Future Generation Computer Systems*, 101:534–541, 2019. 3

[16] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018. 97

[17] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2212–2220, 2019. 97

[18] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. Memetically optimized MCWLD for matching sketches with digital face images. *IEEE Transactions on Information Forensics and Security*, 7(5):1522–1535, 2012. XIII, 5, 7, 8, 20, 30, 31, 33, 35, 43, 52, 53, 60

[19] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008. 24

[20] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016. 74, 99

[21] A. Bosch, A. Zisserman, and X. Muñoz. Image classification using random forests and ferns. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007. 32

[22] M. B. Brewer. In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological bulletin*, 86(2):307, 1979. 16, 73

[23] J. Bryner. This supercomputer can calculate in 1 second what would take you 6 billion years. https://www.livescience.com/62827-fastest-supercomputer.html. 1

[24] J. Buolamwini and T. Gebru. Gender Shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, volume 81, pages 77–91, 2018. 9, 14, 74, 118, 119, 121

[25] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 67–74, 2018. 73, 77, 105, 120

[26] Y. Cao, C. Wang, L. Zhang, and L. Zhang. Edgel index for large-scale sketch-based image search. In *International Conference on Computer Vision and Pattern Recognition*, pages 761–768, 2011. 42

[27] D. Card, M. Zhang, and N. A. Smith. Deep weighted averaging classifiers. In *Conference on Fairness, Accountability, and Transparency*, pages 369–378, 2019. 74

[28] M. Carlsson and S. Eriksson. In-group gender bias in hiring: Real-world evidence. *Economics Letters*, 185:108686, 2019. 73

[29] C.-H. Chan and J. Kittler. Angular sparsemax for face recognition. In *International Conference on Pattern Recognition*, pages 10473–10479. IEEE, 2021. 4

[30] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, 2015. 79

[31] P. Chiroro and T. Valentine. An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 48(4):879–894, 1995. 73

[32] J. Choi, A. Sharma, D. W. Jacobs, and L. S. Davis. Data insufficiency in sketch versus photo face recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2012. 21

[33] T. Chugh, H. S. Bhatt, R. Singh, and M. Vatsa. Matching age separated composite sketches and digital face images. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, 2013. 20

[34] T. Chugh, M. Singh, S. Nagpal, R. Singh, and M. Vatsa. Transfer learning based evolutionary algorithm for composite face sketch recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 30

[35] A. Coston, K. N. Ramamurthy, D. Wei, K. R. Varshney, S. Speakman, Z. Mustahsan, and S. Chakraborty. Fair transfer learning with missing protected attributes. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, Honolulu, HI, USA*, 2019. 97

[36] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pages 1436–1445, 2019. 99

[37] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 53, 67

[38] S. Damas, O. Cordón, O. Ibáñez, J. Santamaría, I. Alemán, M. Botella, and F. Navarro. Forensic identification by computer-aided craniofacial superimposition: A survey. *ACM Computing Surveys*, 43(4):27:1–27:27, 2011. 65

[39] A. Das, A. Dantcheva, and F. Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *European Conference of Computer Vision Workshops*, 2018. 14, 74, 99

[40] T. de Freitas Pereira, A. Anjos, and S. Marcel. Heterogeneous face recognition using domain specific units. *IEEE Transactions on Information Forensics and Security*, 14(7):1803–1816, 2018. 43

[41] T. de Vries, I. Misra, C. Wang, and L. van der Maaten. Does object recognition work for everyone? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019. 13

[42] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 128

[43] S. Dey, P. Riba, A. Dutta, J. Llados, and Y.-Z. Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2179–2188, 2019. 42

[44] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020. 111

[45] M. Du, F. Yang, N. Zou, and X. Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, pages 1–1, 2020. 111

[46] F. Duan, Y. Yang, Y. Li, Y. Tian, K. Lu, Z. Wu, and M. Zhou. Skull identification via correlation measure between skull and face shape. *IEEE Transactions on Information Forensics and Security*, 9(8):1322–1332, 2014. 65

[47] A. Dutta and Z. Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2019. 40, 42

[48] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018. 99

[49] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014. 79

[50] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE Transactions on Visualization and Computer Graphics*, 17(11):1624–1636, 2011. 42

[51] H. D. Ellis, J. B. Deregowski, and J. W. Shepherd. Descriptions of white and black faces by white and black subjects. *International Journal of Psychology*, 10(2):119–123, 1975. 83, 84

[52] Y. Fang, W. Deng, J. Du, and J. Hu. Identity-aware cyclegan for face photo-sketch synthesis and recognition. *Pattern Recognition*, 102, 2020. 42

[53] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Conference on Fairness, Accountability, and Transparency*, pages 329–338, 2019. 73

[54] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He. Dvg-face: Dual variational generation for heterogeneous face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 20, 43

[55] S. Fu, H. He, and Z.-G. Hou. Learning race from face: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2483–2509, 2014. 80

[56] J. Gaston, J. Ming, and D. Crookes. Matching larger image areas for unconstrained face identification. *IEEE Transactions on Cybernetics*, 49(8):3191–3202, 2019. 97

[57] S. Ge, S. Zhao, X. Gao, and J. Li. Fewer-shots and lower-resolutions: Towards ultrafast face recognition in the wild. In *ACM International Conference on Multimedia*, pages 229–237, 10 2019. 74

[58] S. C. Geyik, S. Ambler, and K. Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2221–2231, 2019. 97

[59] B. Glymour and J. Herington. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Conference on Fairness, Accountability, and Transparency*, pages 269–278, 2019. 74, 97

[60] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 13

[61] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010. 4

[62] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102, 2016. 73, 77

[63] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1735–1742, 2006. 48

[64] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain. Matching composite sketches to face photos: A Component-Based Approach. *IEEE Transactions on Information Forensics and Security*, 8(1):191–204, 2013. 8, 20, 43

[65] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 26, 53, 72, 77, 105

[66] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811, 2018. 9, 14, 74

[67] P. J. Hills and M. B. Lewis. Reducing the own-race bias in face recognition by shifting attention. *The Quarterly Journal of Experimental Psychology*, 59(6):996–1002, 2006. 83

[68] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 72, 77

[69] R. Hu and J. Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, 117(7):790 – 806, 2013. 42

[70] D. Huang and Y. F. Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *IEEE International Conference on Computer Vision*, pages 2496–2503, 2013. 43

[71] F. Huang, C. Jin, Y. Zhang, K. Weng, T. Zhang, and W. Fan. Sketch-based image retrieval with deep visual semantic descriptor. *Pattern Recognition*, 76:537 – 548, 2018. 42

[72] M. I. Huete, O. Ibáñez, C. Wilkinson, and T. Kahana. Past, present, and future of craniofacial superimposition: Literature and international surveys. *Legal Medicine*, 17(4):267–278, 2015. 65

[73] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 60

[74] S. Jannati, A. Kumar, A. Niessen-Ruenzi, and J. Wolfers. In-group bias in financial markets. *Available at SSRN 2884218*, 2020. 73

[75] J. Jiang, Y. Yu, Z. Wang, X. Liu, and J. Ma. Graph-regularized locality-constrained joint dictionary and residual learning for face sketch synthesis. *IEEE Transactions on Image Processing*, 28(2):628–641, 2019. 60

[76] S. Kapur. Reducing racial bias in ai models for clinical use requires a top-down intervention. *Nature Machine Intelligence*, 3(6):460–460, 2021. 97

[77] K. Kärkkäinen and J. Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv preprint arXiv:1908.04913*, 2019. 98, 103, 104

[78] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016. 4, 19, 88

[79] R. Keshari, R. Singh, and M. Vatsa. Guided dropout. In *AAAI Conference on Artificial Intelligence*, pages 4065–4072, 2019. 100

[80] Z. Khan and Y. Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 587–597, 2021. 97

[81] Z. Khan, Y. Hu, and A. Mian. Facial self similarity for sketch to photo matching. In *International Conference on Digital Image Computing Techniques and Applications*, pages 1–7, 2012. 20

[82] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim. Learning not to learn: Training deep neural networks with biased data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019. 99, 130

[83] B. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 35(6):1410–1422, 2013. 20, 43

[84] B. F. Klare, S. S. Bucak, A. K. Jain, and T. Akgul. Towards automated caricature recognition. In *IAPR International Conference on Biometrics*, pages 139–146, 2012. 43, 51, 52, 56

[85] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. J. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, 2015. 4, 19

[86] B. F. Klare, Z. Li, and A. Jain. Matching forensic sketches to mug shot photos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):639–646, 2011. 21, 39

[87] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain. The facesketchid system: Matching facial composites to mugshots. *IEEE Transactions on Information Forensics and Security*, 9(12):2248–2263, 2014. 20, 21, 30, 31

[88] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 9

[89] K. S. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020. 111

[90] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105. 2012. XIII, 9, 26, 53, 121

[91] B. Kulis, M. Sustik, and I. Dhillon. Learning low-rank kernel matrices. In *International Conference on Machine Learning*, pages 505–512, 2006. 23

[92] A. Lanitis. Comparative evaluation of automatic age-progression methodologies. *EURASIP Journal on Advances in Signal Processing*, pages 101:1–101:10, 2008. 30

[93] S. Lebrecht, L. J. Pierce, M. J. Tarr, and J. W. Tanaka. Perceptual other-race training reduces implicit racial bias. *PloS one*, 4(1):e4215, 2009. 16

[94] S. Lebrecht, L. J. Pierce, M. J. Tarr, and J. W. Tanaka. Perceptual other-race training reduces implicit racial bias. *PloS one*, 4(1):e4215, 2009. 85

[95] D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999. 22, 32

[96] Y. Li and N. Vasconcelos. Repair: Removing representation bias by dataset resampling. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 111

[97] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang. Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1089–1102, 2017. 8, 20, 21, 33, 35, 56, 60

[98] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 100

[99] D. Liu, X. Gao, N. Wang, C. Peng, and J. Li. Iterative local re-ranking with attribute guided synthesis for face sketch recognition. *Pattern Recognition*, 109:107579, 2021. 43

[100] P. Liu, Y. Lin, Z. Meng, L. Lu, W. Deng, J. T. Zhou, and Y. Yang. Point adversarial self-mining: A simple method for facial expression recognition. *IEEE Transactions on Cybernetics*, 2021. 97

[101] Q. Liu, L. Xie, H. Wang, and A. L. Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *IEEE International Conference on Computer Vision*, pages 3662–3671, 2019. 42

[102] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 2015. 4

[103] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9:2579–2605, 2008. 89

[104] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393, 2018. 99

[105] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Conference on Fairness, Accountability, and Transparency*, pages 349–358, 2019. 74

[106] A. Majumdar, R. Singh, and M. Vatsa. Face verification via class sparsity based supervised encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1273–1280, 2017. 32

[107] D. Mandal and S. Biswas. Generalized coupled dictionary learning approach with applications to cross-modal matching. *IEEE Transactions on Image Processing*, 25(8):3826–3837, 2016. 60

[108] V. Manjunatha, N. Saini, and L. S. Davis. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9562–9571, 2019. 14

[109] A. Martínez and R. Benavente. The AR face database. Technical Report 24, Computer Vision Center, 1998. XIII, 5, 31

[110] I. Masi, F.-J. Chang, J. Choi, S. Harel, J. Kim, K. Kim, J. Leksut, S. Rawls, Y. Wu, T. Hassner, et al. Learning pose-aware models for pose-invariant face recognition in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):379–393, 2018. 4

[111] R. Mauro and M. Kubovy. Caricature and face recognition. *Memory and Cognition*, 20(4):433–440, 1992. 39

[112] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 2021. 4

[113] C. A. Meissner and J. C. Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1):3–35, 2001. 71, 80

[114] A. Mignon and F. Jurie. CMML: A New Metric Learning Approach for Cross Modal Matching. In *Asian Conference on Computer Vision*, page 14, 2012. 20

[115] A. Mishra, S. N. Rai, A. Mishra, and C. V. Jawahar. IIIT-CFW: A benchmark database of cartoon faces in the wild. In *European Conference on Computer Vision Workshops*, pages 35–47, 2016. 51, 52

[116] P. Mittal, A. Jain, G. Goswami, M. Vatsa, and R. Singh. Composite sketch recognition using saliency and attribute feedback. *Information Fusion*, 33:86–99, 2017. 8, 20

[117] P. Mittal, M. Vatsa, and R. Singh. Composite sketch recognition via deep network - a transfer learning approach. In *International Conference on Biometrics*, pages 251–256, 2015. 8, 20, 21, 31

[118] S. Nagpal, M. Singh, A. Jain, R. Singh, M. Vatsa, and A. Noore. On matching skulls to digital face images: A preliminary approach. In *IEEE International Joint Conference on Biometrics*, pages 813–819, 2017. XV, 65, 66, 67, 68

[119] S. Nagpal, M. Singh, R. Singh, and M. Vatsa. Deep learning for face recognition: Pride or prejudiced?, 2019. arXiv,1904.01219. 9, 97

[120] S. Nagpal, M. Singh, R. Singh, and M. Vatsa. Attribute aware filter-drop for bias-invariant classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. XVII, XVIII, 74, 99, 100, 109, 119, 122, 125

[121] S. Nagpal, M. Singh, R. Singh, and M. Vatsa. Diversity blocks for de-biasing classification models. In *IEEE International Joint Conference on Biometrics*, pages 1–9, 2020. XIV, 10, 11, 13, 136

[122] S. Nagpal, M. Singh, R. Singh, and M. Vatsa. Discriminative shared transform learning for sketch to image matching. *Pattern Recognition*, 114:107815, 2021. XIV, 47, 68

[123] S. Nagpal, M. Singh, R. Singh, M. Vatsa, A. Noore, and A. Majumdar. Face sketch matching via coupled deep transform learning. In *IEEE International Conference on Computer Vision*, pages 5429–5438, 2017. 37, 39, 45, 53, 56, 58, 60, 64

[124] S. Nagpal, M. Vatsa, and R. Singh. Sketch recognition: What lies ahead? *Image and Vision Computing*, 55, Part 1:9 – 13, 2016. 21, 65

[125] M. E. Nicholls, O. Churches, and T. Loetscher. Perception of an ambiguous figure is affected by own-age social biases. *Scientific reports*, 8(1), 2018. 71

[126] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 53

[127] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311 – 3325, 1997. 22

[128] L. Oneto, M. Doninini, A. Elders, and M. Pontil. Taking advantage of multitask learning for fair classification. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 227–237, 2019. 97

[129] S. Ouyang, T. Hospedales, Y. Song, X. Li, C. C. Loy, and X. Wang. A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution. *Image and Vision Computing*, 56:28 – 48, 2016. 39, 42, 65

[130] S. Ouyang, T. M. Hospedales, Y. Z. Song, and X. Li. Forgetmenot: Memory-aware forensic facial sketch matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5571–5579, 2016. 20, 21, 43

[131] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *IEEE international conference on multimedia and Expo*, pages 5–pp, 2005. XIII, 5

[132] S. Park, J. Park, S.-J. Shin, and I.-C. Moon. Adversarial dropout for supervised and semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, 2018. 100

[133] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 32, 53, 121

[134] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conference on Signals, Systems and Computers*, pages 1–3, 1993. 23

[135] G. K. Patro, A. Biswas, N. Ganguly, K. P. Gummadi, and A. Chakraborty. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *The Web Conference*, WWW '20, page 1194–1204. Association for Computing Machinery, 2020. 97

[136] C. Peng, X. Gao, N. Wang, and J. Li. Face recognition from multiple stylistic sketches: Scenarios, datasets, and evaluation. *Pattern Recognition*, 84:262 – 272, 2018. 42

[137] C. Peng, N. Wang, J. Li, and X. Gao. Universal face photo-sketch style transfer via multiview domain translation. *IEEE Transactions on Image Processing*, 29:8519–8534, 2020. 60

[138] L. A. Pereira and R. da Silva Torres. Semi-supervised transfer subspace for domain adaptation. *Pattern Recognition*, 75:235 – 249, 2018. 39

[139] L. Pfister and Y. Bresler. Automatic parameter tuning for image denoising with learned sparsifying transforms. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6040–6044, 2017. 24

[140] P. J. Phillips, Hyeonjoon Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000. 53

[141] A. Puc, V. Štruc, and K. Grm. Analysis of race and gender bias in deep age estimation models. In *European Signal Processing Conference*, pages 830–834. IEEE, 2021. 97

[142] Y. Qi, Y. Song, H. Zhang, and J. Liu. Sketch-based image retrieval via siamese convolutional neural network. In *IEEE International Conference on Image Processing*, pages 2460–2464, 2016. 42

[143] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Conference on Fairness, Accountability, and Transparency*, page 469–481, 2020. 9, 74, 97

[144] I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *AAAI/ACM Conference on AI Ethics and Society*, 2019. 69

[145] C. Rathgeb, D. Dogan, F. Stockhardt, M. De Marsico, and C. Busch. Plastic surgery: An obstacle for deep face recognition? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–807, 2020. 4

[146] S. Ravishankar and Y. Bresler. Learning sparsifying transforms. *IEEE Transactions on Signal Processing*, 61(5):1072–1086, 2013. 22, 23, 28, 54

[147] S. Ravishankar and Y. Bresler. Efficient blind compressed sensing using sparsifying transforms with convergence guarantees and application to magnetic resonance imaging. *SIAM Journal on Imaging Sciences*, 8(4):2519–2557, 2015. 23

[148] S. Ravishankar and Y. Bresler. Online sparsifying transform learning 2014; Part II: Convergence Analysis. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):637–646, 2015. 23, 49

[149] S. Ravishankar and Y. Bresler. Data-driven learning of a union of sparsifying transforms model for blind compressed sensing. *IEEE Transactions on Computational Imaging*, 2(3):294–309, 2016. 24

[150] S. Ravishankar, B. Wen, and Y. Bresler. Online sparsifying transform learning 2014; part I: Algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):625–636, 2015. 23, 24

[151] A. W. Rawls and K. Ricanek. MORPH: Development and optimization of a longitudinal age progression database. In *Biometric ID Management and Multimodal Communication*, pages 17–24, 2009. XIII, 5, 78

[152] P. Reavy and L. Egan. Skull remains unidentified 3 years after its discovery. https://www.ksl.com/?nid=148&sid=44886383. Posted: 3rd July, 2017. 64

[153] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner. Face recognition: too bias, or not too bias? *arXiv preprint arXiv:2002.06483*, 2020. 13, 74

[154] R. Rothe, R. Timofte, and L. V. Gool. DEX: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops*, pages 10–15, 2015. 79

[155] H. J. Ryu, H. Adam, and M. Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2018. 14, 74

[156] A. Selvraj. Police sketch: Not a picture perfect tool to crack all criminal cases. https://timesofindia.indiatimes.com/blogs/tracking-indian-communities/police-sketch-not-a-picture-perfect-tool-to-crack-all-criminal-cases 5

[157] L. Sheffer. Partisan in-group bias before and after elections. *Electoral Studies*, 67:102191, 2020. 16, 73

[158] Y. Shi, D. Deb, and A. K. Jain. Warpgan: Automatic caricature generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10762–10771, 2019. 43

[159] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003. XIII, 4, 5, 78

[160] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 26

[161] M. Singh, S. Nagpal, R. Singh, M. Vatsa, and A. Noore. Learning a shared transform model for skull to digital face image matching. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, 2018. 41, 67

[162] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006. 83

[163] J. Song, Y.-Z. Song, T. Xiang, T. Hospedales, and X. Ruan. Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval. In *Proceedings of the British Machine Vision Conference*, pages 132.1–132.11, 2016. 58

[164] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *International Conference on Computer Vision*, pages 5552–5561, 2017. 40, 42, 58

[165] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2239–2248, 2018. 97

[166] M. Srivastava, H. Heidari, and A. Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2459–2468, 2019. 13

[167] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 100

[168] P. Stock and M. Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *European Conference on Computer Vision*, pages 498–512, 2018. 9

[169] K. Takayama, H. Johan, and T. Nishita. Face detection and face recognition of cartoon characters using feature extraction. In *Image, Electronics and Visual Computing Workshop*, page 48, 2012. 43

[170] Y. H. Tsai, H. M. Hsu, C. A. Hou, and Y. C. F. Wang. Person-specific domain adaptation with applications to heterogeneous face recognition. In *IEEE International Conference on Image Processing*, pages 338–342, 2014. 20

[171] P. Tu, R. Book, X. Liu, N. Krahnstoever, C. Adrian, and P. Williams. Automatic face recognition from skeletal remains. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007. 65

[172] R. G. Uhl and N. da Vitoria Lobo. A framework for recognizing a facial image from a police sketch. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–593, 1996. 8

[173] S. Vats, S. Jain, and P. Guha. A novel ensemble framework for face search. In *International Conference on Pattern Recognition*, pages 514–528. Springer, 2021. 39

[174] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 67

[175] A. Wang, J. Cai, J. Lu, and T.-J. Cham. MMSS: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *The IEEE International Conference on Computer Vision*, 2015. 32, 35, 56

[176] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 128

[177] H. Wang, X. Dong, Z. Jin, J.-L. Dugelay, and M. Tistarelli. Cross-spectrum face recognition using subspace projection hashing. In *International Conference on Pattern Recognition*, pages 615–622. IEEE, 2021. 39

[178] H. Wang, S. Wang, Z. Jin, Y. Wang, C. Chen, and M. Tistarelli. Similarity-based gray-box adversarial attack against deep face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8, 2021. 4

[179] L. Wang, V. Sindagi, and V. M. Patel. High-quality facial photo-sketch synthesis using multi-adversarial networks. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 83–90, 2018. 53, 60

[180] M. Wang and W. Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 99

[181] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *IEEE International Conference on Computer Vision*, pages 692–702, 2019. 74, 78

[182] S. Wang, Z. Ding, and Y. Fu. Coupled marginalized auto-encoders for cross-domain multi-view learning. In *International Joint Conferences on Artificial Intelligence*, pages 2125–2131, 2016. 39

[183] S. Wang, L. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2216–2223, 2012. 8, 32, 35, 60

[184] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *IEEE International Conference on Computer Vision*, 2019. 74, 99

[185] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009. XIII, 7, 8, 21, 30, 31, 43, 52, 53, 60

[186] Y. Wang, F. Huang, Y. Zhang, R. Feng, T. Zhang, and W. Fan. Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval. *Pattern Recognition*, 100, 2020. 39

[187] Y. Wang, Y. Y. Tang, L. Li, and H. Chen. Modal regression-based atomic representation for robust face recognition and reconstruction. *IEEE Transactions on Cybernetics*, 50(10):4393–4405, 2020. 97

[188] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 99

[189] H. Wiese, J. Komes, and S. R. Schweinberger. Ageing faces in ageing minds: a review on the own-age bias in face recognition. *Visual Cognition*, 21(9-10):1337–1363, 2013. 88

[190] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE International Conference on Computer Vision*, pages 529–534. IEEE, 2011. 4

[191] L. Wolf, Y. Taigman, and A. Polyak. Unsupervised creation of parameterized avatars. In *IEEE International Conference on Computer Vision*, pages 1539–1547, 2017. 43, 54

[192] X. Wu, R. He, Z. Sun, and T. Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 11, 32, 72, 77, 120

[193] X. Wu, L. Song, R. He, and T. Tan. Coupled deep learning for heterogeneous face recognition. In *AAAI Conference on Artificial Intelligence*, 2018. 40

[194] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016. 100

[195] S. Xie, H. Hu, and Y. Wu. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognition*, 92:177–191, 2019. 84

[196] C. Xinyuan, W. Chunheng, X. Baihua, C. Xue, L. Zhijian, and S. Yanqin. Coupled latent least squares regression for heterogeneous face recognition. In *IEEE International Conference on Image Processing*, pages 2772–2776, 2013. 20

[197] D. Xu, J. Song, X. Alameda-Pineda, E. Ricci, and N. Sebe. Multi-paced dictionary learning for cross-domain retrieval and recognition. In *International Conference on Pattern Recognition*, pages 3228–3233, 2016. 60

[198] D. Yadav, R. Singh, M. Vatsa, and A. Noore. Recognizing age-separated face images: Humans and machines. *PLoS ONE*, 9, 2014. 30

[199] B. Yang, A. J. Ma, and P. C. Yuen. Learning domain-shared group-sparse representation for unsupervised domain adaptation. *Pattern Recognition*, 81:615 – 632, 2018. 39

[200] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 4, 77

[201] Z. Yi, H. R. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE International Conference on Computer Vision*, pages 2868–2876, 2017. 60

[202] J. Yu, X. Hao, H. Xie, and Y. Yu. Fair face recognition using data balancing, enhancement and fusion. In *European Conference on Computer Vision*, pages 492–505, 2020. 99

[203] Q. Yu, F. Liu, Y. SonG, T. Xiang, T. Hospedales, and C. C. Loy. Sketch me that shoe. In *International Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016. 42, 52, 58

[204] S. Yucer, S. Akcay, N. Al-Moubayed, and T. P. Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 99

[205] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017. 99, 130

[206] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013. 99

[207] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 97, 99

[208] H. Zhang, P. She, Y. Liu, J. Gan, X. Cao, and H. Foroosh. Learning structural representations via dynamic object landmarks discovery for sketch recognition and retrieval. *IEEE Transactions on Image Processing*, 28(9):4486–4499, 2019. 42

[209] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 513–520, 2011. 8, 21, 31, 43, 52, 53, 60

[210] X. Zhang, Y. Huang, Q. Zou, Y. Pei, R. Zhang, and S. Wang. A hybrid convolutional neural network for sketch recognition. *Pattern Recognition Letters*, 130:73–82, 2020. 42

[211] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, 2017. XIV, XVII, 10, 98, 103, 104, 112, 118, 121

[212] W. Zheng, L. Yan, C. Gou, W. Zhang, and F. Wang. A relation network embedded with prior features for few-shot caricature recognition. In *IEEE International Conference on Multimedia and Expo*, pages 1510–1515, 2019. 57

[213] W. Zheng, L. Yan, F.-Y. Wang, and C. Gou. Learning from the past: Meta-continual learning with knowledge embedding for jointly sketch, cartoon, and caricature face recognition. In *ACM International Conference on Multimedia*, pages 736–743, 2020. 20, 43

[214] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 77

[215] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2242–2251, 2017. 60

[216] J. Zou and L. Schiebinger. AI can be sexist and racist — it's time to make it fair , 2018. https://www.nature.com/articles/d41586-018-05707-8. 9