INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY DELHI

# Facial Analysis under Low Resolution and Disguise Variations

by

Maneet Singh

Under the supervision of

Dr. Richa Singh

Dr. Mayank Vatsa

Indraprastha Institute of Information Technology Delhi

December, 2021

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY DELHI

# Facial Analysis under Low Resolution and Disguise Variations

by

Maneet Singh

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

to the
Indraprastha Institute of Information Technology Delhi
December, 2021

# Certificate

This is to certify that the thesis titled "**Facial Analysis under Low Resolution and Disguise Variations**" being submitted by **Maneet Singh** 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 him 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.

December, 2021
Dr. Richa Singh
Professor

December, 2021
Dr. Mayank Vatsa
Professor

Indraprastha Institute of Information Technology Delhi
New Delhi

# Acknowledgment


> I've learned that people will forget what you said, people will forget what you did, but people will never forget how you made them feel.
>
> Maya Angelou

This dissertation marks the end of a crucial chapter of my life, and I was fortunate to have the support of several fantastic people during its course. Each interaction taught me something unique and shaped me into the person I am today.

I would like to begin by expressing my heartfelt gratitude to my advisors, **Prof. Richa Singh** and **Prof. Mayank Vatsa**, for their trust, support, and continuous boost for achieving bigger heights. It is through them that I have learned the value of persistence and self-belief; crucial qualities that will support me throughout my life and career. I would also like to thank **Prof. Afzel Noore** for hosting me during my internship at the West Virginia University. His compassion and kindness continues to encourage me to be a better human being, while his technical expertise and critical thinking inspire me to develop into a better researcher as well. My sincere gratitude to **Prof. Rama Chellappa**, **Dr. Nalini Ratha**, **Dr. Arun Ross**, and **Dr. Angshul Majumdar**; each of whom have contributed towards the improvement of my technical writing and critical thought process via collaboration on different research papers.

I would like to express my sincere gratitude to **IIIT-Delhi**, which has been like my home for the past several years. The support provided via the Institute in terms of the infrastructure, facilities, and helping staff remains at par with some of the best that I have seen till date. I must also express my gratitude to **Dr. Pankaj Jalote** for enabling me to roll-over from the B.Tech program to the Ph.D. program and fostering a conducive research environment at the Institute. A huge shout-out to **Priti Ma'am** for being the sole point of contact for all my administrative dilemmas and for resolving them with a warm smile. I would also like to acknowledge the **Department of Electronics and Information Technology, Government of India**, for their financial support during my Ph.D. by funding my research. A special mention to **Dr. Daksha** and **Dr. Naman**, who not only made my internship at WVU memorable, but supported me beyond that engagement with

their surprisingly sharp memory and little gestures. I am also grateful to my collaborators: **Shruti, Soumyadeep, Akshay Agarwal, Akshay Sethi, Mohit Chawla, Vineet, Amlaan, Prateekshit, Isha, Arushi, Sanchit Gupta, Sanchit Sinha, Nikita,** and **Mohit Agarwal**. Each one of you taught me something unique at the personal and professional front.

Finally, the past few years would have been impossible without the unconditional support of my family and friends. My parents - **Dr. Kawaldeep** and **Amarjeet**, who gave me the freedom to pursue the path I wanted to, and taught me that nothing is impossible if one has belief in themselves. You taught me to take everything with a pinch of laughter, and that has been the guiding light in this roller coaster journey. My sister, **Guneet** and brother-in-law, **Lokesh**, who have supported me through the thick and thin, and have been my confidant across years. **Neeta aunty** - who has never failed to surprise me with her selflessness and subtle ways of showing care by always pampering me with food. My friends since forever - **Jayasi, Vidushi,** and **Shreya**, with whom I've had several sessions about almost everything under the sun. This journey would have been incomplete without either of your support. And finally, **Shruti**, who has been *my person* throughout this journey, and continues to remain so. I could not have asked for a better research partner, collaborator, brain-stormer, and support system. This dissertation would not have been possible without you being there during the best and worst of times. Each person mentioned above has touched my life and contributed towards the fulfillment of this work. I hope this dissertation does you all proud.

**Facial Analysis under Low Resolution and Disguise Variations**

by

Maneet Singh

## Abstract

Automated facial analysis has widespread applicability in scenarios related to image tagging, access control, and surveillance. Initial research focused primarily on face recognition in constrained settings, where the captured face image had variations due to pose, illumination, or expression. With the increased applicability of facial analysis models in real world scenarios, dedicated research was required for data captured in unconstrained settings including resolution variations. When subjects are captured at a large stand-off distance from the acquisition device, the resulting region of interest (ROI) capturing the face is often small (less than $32 \times 32$), requiring recognition of low resolution or very low resolution facial regions. Data captured in such unconstrained scenarios also often contain people using different disguise accessories or occluded faces, resulting in the obfuscation of the face region, rendering automated face recognition challenging. To this effect, this dissertation focuses on facial analysis under low resolution and disguise variations. Two facial recognition algorithms have been presented for data captured in low and very resolution settings: Dual Directed Capsule Network and DeriveNet model, followed by two novel datasets (Disguised Faces in the Wild 2018 and 2019) for facilitating research on disguised faces in the wild along with a Disguise-Resilient face verification framework. This is followed by designing facial analysis models for attribute prediction in low and very low resolution settings.

We begin with developing deep learning algorithms for low or very low resolution face recognition, which suffers from the challenge of limited interpretable information in the face images, thus resulting in ineffective feature extraction and classification. In order to address this challenge, we propose two novel algorithms: Dual Directed Capsule Network (DirectCapsNet) and DeriveNet model. Since low resolution face images contain limited meaningful information, we propose utilizing a small set of high resolution samples for directing the classification model towards learning richer features. The DirectCapsNet is built using a combination of convolutional and capsule layers, and is trained via three loss functions: HR-Anchor loss, Reconstruction loss, and Margin loss. DeriveNet thus learns rich feature representations for very low resolution samples by utilizing the auxiliary high resolution samples during training. While capsule layers encode rich features, they are computationally expensive and contain a larger number of trainable parameters. In order to address the above limitation, a novel DeriveNet model has been proposed for low and very low resolution face recognition. The proposed model utilizes a set of high resolution images for learning an effective recognition model via combination of two loss functions: Derived-Margin softmax loss and Reconstruction-Center loss. The proposed Derived-Margin softmax loss estimates the inter-class variations between low resolution samples and models it as a margin for learning improved classification boundaries. Experimental analysis is performed on different challenging real-world datasets including the UnConstrained College Students (UCCS)

dataset for facial regions having less than $24 \times 24$ resolution. Comparison with recent techniques demonstrates state-of-the-art results by the proposed algorithm.

The next contribution of this dissertation lies in the area of disguised face recognition, where individuals attempt to obfuscate the face region, either intentionally in order to fool the automated system, or unintentionally by the use of day-to-day facial accessories. To the best of our knowledge, most of the research focused on disguised face recognition in constrained scenarios, with limited disguise accessories and other variations. Therefore, as part of this dissertation, we propose the Disguised Faces in the Wild (DFW) 2018 and DFW2019 datasets containing face images with unconstrained disguise variations, captured across different resolutions, acquisition devices, lighting, pose, and expression. The datasets were released as part of two international workshops for facilitating research in this direction. We also present the Disguise-Resilient framework using a novel Disguise Encoder-Decoder network, with application to face verification. The efficacy of the proposed framework has been demonstrated on the challenging DFW2018 and DFW2019 datasets, where it achieves state-of-the-art performance. Further, the arduous task of disguised face recognition in low resolution settings has also been explored and presented to the research community. Baseline results and performance of the proposed framework for face images with resolutions varying from $32 \times 32$ to $16 \times 16$ demand dedicated research focus from the community.

The final contribution of this dissertation focuses on developing deep learning algorithms for learning discriminative features, with application to attribute classification in low resolution face images. Automated prediction of attributes such as gender (male/female) or adulthood (adult/child) can be useful as ancillary information for person identification, enhanced human computer interaction, or for restricting age-based access. As part of this contribution, two supervised variations of the deep learning based Autoencoder model are proposed for learning class-specific features: Class Specific Mean Autoencoder and Class Representative Autoencoder. Both models utilize the concept that the mean feature of a given class contains class-specific information which can be incorporated for learning discriminative rich features. To the best of our knowledge, this is one of the initial research focused on analyzing attributes in low resolution facial regions. The proposed autoencoder models are able to extract meaningful information by modeling the inter-class and intra-class variations, resulting in improved performance for low resolution attribute classification from face images. Experimental evaluation on different datasets for facial images of $24 \times 24$ and $16 \times 16$ resolution demonstrate the effectiveness of the techniques.

# Table of Contents


**1 Introduction** **1**

1.1 (Very) Low Resolution Facial Analysis . . . . . 5

1.2 Disguised Face Recognition . . . . . 9

1.3 Research Contributions . . . . . 12

**2 Dual Directed Capsule Network for Very Low Resolution Image Recognition** **17**

2.1 Introduction . . . . . 17

2.2 Related Work . . . . . 19

2.3 Proposed Dual Directed Capsule Network . . . . . 21

2.3.1 Preliminaries: Capsule Networks . . . . . 21

2.3.2 Proposed DirectCapsNet . . . . . 23

2.3.3 Implementation Details . . . . . 26

2.4 Experiments and Protocols . . . . . 27

2.5 Results and Analysis . . . . . 28

2.6 Summary . . . . . 33

**3 DeriveNet for (Very) Low Resolution Image Classification** **35**

3.1 Introduction . . . . . 35

3.2 Related Work . . . . . 37

3.3 DeriveNet for (V)LR Recognition . . . . . 39

3.3.1 Derived-Margin Softmax Loss ($\mathcal{L}_{D-Margin}$) . . . . . 40

3.3.2 Reconstruction-Center Loss ($\mathcal{L}_{ReCent}$) . . . . . 42

3.3.3 Training DeriveNet with Multi-Resolution Pyramid based Data Augmentation . . . 44
3.3.4 Implementation Details . . . 44
3.4 Datasets and Experimental Protocol . . . 45
3.5 Results and Analysis . . . 47
3.6 Summary . . . 55

**4 Disguised Faces in the Wild** **57**
4.1 Introduction . . . 57
4.2 Motivation . . . 59
4.3 Disguised Faces in the Wild (DFW) 2018 Dataset . . . 62
4.3.1 Dataset Statistics . . . 62
4.3.2 Protocols for Evaluation . . . 64
4.3.3 Nomenclature and Data Distribution . . . 65
4.4 Disguised Faces in the Wild 2018 Competition . . . 66
4.4.1 Baseline Results . . . 67
4.4.2 DFW2018 Competition: Submissions . . . 67
4.4.3 Results . . . 72
4.5 DFW2018 Dataset: Easy, Medium, and Hard Degree of Difficulty . . . 76
4.6 Disguised Faces in the Wild 2019 Dataset . . . 79
4.6.1 Protocols for Evaluation . . . 81
4.7 Baseline Results . . . 82
4.8 Disguised Faces in the Wild 2019 Competition . . . 83
4.8.1 DFW2019 Competition: Submissions . . . 83
4.8.2 Results . . . 85
4.9 DFW2019 Dataset: Easy, Medium, and Hard Pairs . . . 90
4.10 Summary . . . 93

**5 Disguised Resilient Face Verification** **95**
5.1 Introduction . . . 95
5.2 Related Work . . . 97

5.3 Proposed Disguise Resilient (D-Res) Framework . . . 98
5.3.1 Disguise Encoder-Decoder Network (DED-Net) . . . 99
5.3.2 D-Res Framework for Face Verification with Disguise Variations . . . 103
5.4 Experiments and Implementation Details . . . 104
5.5 Results and Analysis . . . 106
5.5.1 Performance of the D-Res Framework . . . 107
5.5.2 Baseline Results and Performance of D-Res Framework for Low Resolution Disguised Face Verification . . . 110
5.5.3 Additional Results: Benchmark Face Verification Datasets . . . 112
5.6 Summary . . . 114

**6 Gender Prediction from Very Low Resolution Face Images 117**
6.1 Introduction . . . 117
6.2 Proposed Supervised Autoencoders . . . 119
6.2.1 Preliminaries: Supervised Autoencoders . . . 119
6.2.2 Proposed Class Specific Mean Autoencoder . . . 122
6.2.3 Proposed Class Representative Autoencoder . . . 125
6.3 Datasets and Experimental Protocol . . . 129
6.3.1 Datasets Used . . . 129
6.3.2 Experimental Protocol . . . 129
6.4 Results and Analysis . . . 130
6.5 Additional Experiments on Attribute Prediction . . . 134
6.5.1 Gender Prediction from Low Resolution NIR Images . . . 134
6.5.2 Adulthood Prediction . . . 137
6.6 Summary . . . 147

**7 Conclusion and Future Research 149**

THIS PAGE INTENTIONALLY LEFT BLANK

# List of Figures

1-1 Facial analysis refers to the task of predicting the identity or attribute information (such as age, expression, or gender) from a given face image. . . . . 2

1-2 Sample traditional facial analysis pipeline consisting of different stages. In recent algorithms (such as deep-learning based techniques) some of the steps are combined into one (e.g., a Convolutional Neural Network is used for feature extraction and classification). Often, variations are encountered by a facial analysis pipeline either during image image acquisition (e.g., changes due to pose, resolution, or usage of disguise accessories) or post the image acquisition step (e.g., image/feature tampering or model-based attacks). . . . . 2

1-3 Often in real-world scenarios, the authentication system is developed using facial images captured in constrained environments. However, at real-time the system might encounter images captured in unconstrained scenarios with several covariates such as resolution variations or presence of disguise accessories. The authentication system must therefore be able to match the given input with the constrained gallery samples provided during enrollment. . . . . 3

1-4 Images obtained from the surveillance feed of CCTV cameras related to two unfortunate incidents: (a) Frame captured on a New Delhi metro station depicting two people after abandoning a newborn girl child at the metro station [98], (b) Suspect of the infamous Chennai murder case, where a young woman was murdered at the local railway station [107]. (c) Sample images of the suspect involved in the Boston Marathon Bombing, 2013. All the images suffer from the challenge of low resolution region of interest (facial region), while (c) also suffers from the presence of a disguise accessory (cap). . . . . 4

1-5 Variations observed in facial region at different resolutions: $48 \times 48$, $32 \times 32$, $24 \times 24$, $16 \times 16$, and $8 \times 8$. Reduced resolution results in lesser fine details, low quality, and limited interpretative information content. . . . . 5

1-6 Sample images demonstrating low resolution regions of interest for facial regions. . 6

1-7 Utilization of disguise accessories which obfuscate the facial region often result in large intra-class variations. . . . . 9

1-8 Day-to-day accessories can often result in (a) obscuring different facial regions [136], while such accessories can also be used to (b) impersonate another individual [108], resulting in challenges for an automated face recognition system. . . . . 10

1-9 Key contributions of this thesis focusing on facial analysis under low resolution and disguise variations. . . . . 12

2-1 The proposed DirectCapsNet utilizes HR samples to *direct* learning of more meaningful and discriminative features for VLR image recognition via the proposed HR-anchor loss and the targeted reconstruction loss. . . . . 18

2-2 Sample HR and VLR images from the (a) SVHN dataset, (b) CMU Multi-PIE dataset, and (c) UCCS dataset. The HR images (first row) contain high information content, which is often missing in the VLR samples (second row). . . . . 22

2-3 Architecture of the proposed Dual Directed Capsule Network (DirectCapsNet) for the SVHN dataset [104]. A diagrammatic representation of the HR-Anchor loss is presented for a given class. HR images are used to complement the features learned by the VLR recognition model by directing the model to learn discriminative and information rich features. The figure has been taken from the published manuscript [138]. . . . . 23

2-4 Sample reconstructions obtained on the SVHN dataset from VLR input. DirectCapsNet is able to reconstruct digits where limited information content is available (e.g., green boxes), however it also fails to correctly reconstruct some challenging cases (e.g., red boxes). . . . . 30

2-5 Sample reconstructions obtained from the proposed DirectCapsNet model on the CMU Multi-PIE dataset. For the same class, DirectCapsNet is able to project VLR and HR samples onto a similar target, suggesting robust resolution-invariant feature representations. . . . . 30

2-6 Performance of the proposed DirectCapsNet for varying resolutions of VLR face recognition with pose and illumination variations. The HR resolution was fixed to $36 \times 30$ pixels. Comparison has been shown with HR-LR (MDS) [13] and Mudunuri and Biswas [102]. . . . . 31

2-7 Scores obtained by the proposed DirectCapsNet for VLR recognition on some samples of the UCCS dataset. Each test image has one genuine score (correct class) and 179 imposter scores (incorrect class). . . . . 32

3-1 (a) VLR/LR face images captured in an unconstrained environment often contain limited information content, rendering face recognition challenging. (b) Existing techniques for VLR/LR image classification often focus only on the image, feature, or classification space. The proposed DeriveNet models both the image and classification space for VLR/LR recognition, while implicitly learning effective features. . . . . 36

3-2 Sample HR and VLR/LR images from different datasets: (a)/(b-c) VLR digit/face recognition and (d) VLR/LR face recognition in drone-shot videos. The top and the bottom row contains the HR and VLR/LR images, respectively. . . . . 36

3-3 Illustration of the training phase of the proposed DeriveNet model for VLR face recognition. HR and VLR images are used as input, followed by a convolutional neural network based feature extractor. The extracted embedding are provided as input to two modules: (i) the reconstruction module and the (ii) classification module. During testing, a given VLR input image is passed through the feature extractor, followed by the classification module for obtaining the predicted class. The figure has been taken from the published manuscript [139]. . . . . 38

3-4 Diagrammatic representation of the DeriveNet model for a four class problem, with an input of class-1. The extracted embeddings are provided as input to the reconstruction and classification modules. The distance between the centers of the reconstructions are provided as margins to the proposed Derived-Margin softmax loss for learning a VLR/LR classifier. . . . . 41

3-5 Class Activation Maps obtained using the (i) DeriveNet model and the native (ii) Softmax based model. DeriveNet appears to focus more on the biometric regions of eyes and nose, while the Softmax based model appears to focus more on the hair and other soft features. . . . . 48

3-6 Sample images correctly identified via the proposed DeriveNet model, which are not identified by the other variants used for the ablation study. . . . . 48

3-7 (a) t-SNE plots on the features of (i) Softmax based model and (ii) DeriveNet model. A larger margin is observed between the features learned via DeriveNet. (b) Convergence plot of the DeriveNet model, and (c) Effect of $\lambda$ (Equation (3.1)) on the SVHN dataset. . . . . 50

3-8 Score distributions of the correct and incorrect classes for different resolutions on the UCCS dataset. The Earth Mover's Distance was also calculated between the same class and different class scores for each resolution: 24.55 ($8 \times 8$), 24.65 ($16 \times 16$), 24.72 ($24 \times 24$), and 24.71 ($32 \times 32$). The distance for the smallest resolution ($8 \times 8$) is the least, thus suggesting a larger variation between the scores, as compared to the other resolutions. . . . . 51

3-9 (a-b) CMC curves on the DroneSURF dataset for the two protocols. DeriveNet improves the rank-1 accuracy by over 20% on both protocols. Comparison has also been performed with Amato *et al.* [5]. (c) Sample challenging cases from the DroneSURF dataset which were (i) correctly classified and (ii) mis-classified by the proposed DeriveNet model. . . . . 52

4-1 Authentication systems often face the challenge of matching disguised face images with non-disguised enrolled images. Figure has been taken from the published manuscript [143]. . . . . 58

4-2 Images pertaining to two subjects of the DFW dataset. The dataset contains at most four types of images for each subject: Normal, Validation, Disguised, and Impersonator. . . . . 63

4-3 Diagrammatic representation of (a) AEFRL [148], and (b) UMDNets [7]. Images have been taken from their respective publications. . . . . 69

4-4 Diagrammatic representation of MiRA-Face [187]. Image has directly been taken from their publication. . . . . 69

4-5 ROC curves of all participants along with the baseline results on protocol-1 (impersonation), protocol-2 (obfuscation), and protocol-3 (overall) of the DFW dataset. 71

4-6 Sample False Positive and True Negative pairs reported by a majority of submissions for protocol-1 (impersonation). False Positive refers to the case where an algorithm incorrectly classifies a pair as genuine, and True Negative refers to the case where two samples of different identities are correctly classified as imposters. 74

4-7 Sample False Negative and True Positive pairs reported by a majority of submissions for protocol-2 (obfuscation). False Negative refers to the case where a pair of images are incorrectly classified as an imposter pair, while True Positive refers to the scenario where a pair of images are correctly classified as a genuine pair. . . 75

4-8 Score distribution of the genuine pairs at 0.01% FAR, in terms of three levels of difficulty: easy, medium, and hard. . . . . 78

4-9 Sample *easy* and *hard* pairs of the DFW dataset. . . . . 79

4-10 Sample impersonator pairs created from the IIIT-Delhi Disguise dataset [26]. An individual can often use disguise accessories to impersonate another individual. . . 81

4-11 ROC curves on the DFW2019 dataset for Protocol-1 and Protocol-2. Images have been taken from the published manuscript [136]. . . . . 87

4-12 ROC curves on the DFW2019 dataset for Protocol-3 and Protocol-4. . . . . 89

4-13 Venn diagram demonstrating the number of mis-classifications of the genuine pairs by the top-3 teams (A-4: ArcFaceIntraInter, A-6: FakeFacev2, A-7:Mozart) at 0.01% FAR. The common region (603 samples) is a subset of the *hard* samples which were mis-classified by all algorithms. . . . . 92

4-14 Scatter plot of the scores obtained by the top-3 teams for the Easy, Medium, and Hard pairs of the DFW2019 dataset. . . . . 92

4-15 Venn diagram demonstrating the number of mis-classifications of the genuine pairs (True Positive samples) by the top-3 teams (A-4: ArcFaceIntraInter, A-6: Fake-Facev2, A-7:Mozart) for 0 False Positives. The common region (10,594 samples) corresponds to a subset of samples which were mis-classified by all algorithms. . . 93

5-1 This research presents a novel Disguise Resilient (D-Res) framework for learning disguise invariant features. The framework utilizes the proposed DED-Net architecture for processing tessellated face images. Ten score pairs ($s_i = [c_i, c_2]$) are obtained via the classifier (one for each patch) for genuine and imposter classes. The scores are combined in a weighted manner (using weights $w_i$) to generate the final score $s$ for the given input image pair. . . . . 96

5-2 Overview of the proposed Disguise Resilient Framework on full face images. Pair of images are provided to the DED-Net model for feature extraction, followed by concatenation and input to the classifier. The classifier outputs a score denoting whether the two images belong to the same subject or not. The DED-Net is a convolutional encoder-decoder model, containing convolution (conv.) and deconvolution (deconv.) filters at different layers (represented by squares). It is optimized via the Cosine and Mahalanobis distance based minimization between the input and the reconstruction, along with the Mutual Information based loss between the features. The classifier is a classical neural network containing neurons in each layer (represented as circles). The different colors in the extracted feature signify different values at each position. Figure has been taken from the published manuscript [140]. . . . . 98

5-3 Sample images from the DFW2019 dataset having variations due to (a-b) disguise accessories demonstrating their genuine images (images belonging to the given subject (Subject-A or Subject-B)) and impersonator images (images of different subjects impersonating the given subject). The DFW dataset also contains images having variations due to (c) bridal make-up (sample images of three subjects have been shown before and after applying the make-up). The use of accessories results in obfuscated face regions, rendering automated face recognition challenging. . . . 102

5-4 (a) Histogram of the scores obtained by the D-Res framework on the Overall protocol (DFW2019). Separation can be observed between the genuine and imposter scores with a small overlap. Sample genuine pairs (images belonging to the same subject) are also shown which were not identified by the D-Res framework. The pairs had scores in the overlapping region between the genuine and imposter scores. Excessive make-up and obfuscation results in highly challenging samples. (b) tSNE visualization of the features learned by the D-Res framework for ten subjects suggesting distinguishing features based on the subject information. 108

5-5 Images from the DFW2019 dataset at different resolutions. Images have been bicubically interpolated to $224 \times 224$. . . . . 110

5-6 ROC curves on the (a) DFW2018 and the (b) DFW2019 datasets, for multiple resolutions at the Overall protocol. The performance of the D-Res framework has been compared with the baseline model (LightCNN-29) for different resolutions: (i) Original, (ii) 32x32, (iii) 24x24, and (iv) 16x16. The D-Res framework outperforms the respective baseline at each resolution. . . . . 111

5-7 Bar graphs demonstrating the variation in accuracy obtained over the DFW2018 and DFW2019 datasets for 32x32, 24x24, and 16x16 input resolutions. . . . . 112

6-1 Sample male images from SCface dataset [42] captured from surveillance cameras. 118

6-2 Mean-male and mean-female images obtained from the CMU Multi-PIE dataset [44]. . . . . 119

6-3 Proposed Class Specific Mean Autoencoder. $x$ and $\tilde{x}$ represent the input and the reconstructed samples respectively, $\mathbf{W}_e$ and $\mathbf{W}_d$ denote the encoding and decoding weights, and $f_x$ corresponds to the learned feature vector. . . . . 124

6-4 Pipeline for performing gender recognition on low resolution face images. The proposed AutoGen model is used for feature extraction, followed by a neural network for classification. AutoGen aims to learn discriminative features by incorporating inter-class and intra-class variations during feature learning. Figure has been taken from the published manuscript [137]. . . . . 127

6-5 Receiver Operating Characteristic (ROC) curves for $24 \times 24$ and $16 \times 16$ resolution face images from the CMU Multi-PIE dataset. Here, 'Proposed' refers to the AutoGen model. . . . . 131

6-6 Sample male images misclassified as females by AutoGen. Most of the misclassifications are categorized by unusual hairstyle, expression, or accessories such as sunglasses. . . . . 132

6-7 Comparison of classification accuracies for two layer feature extraction models, for face images having $16 \times 16$ resolution. . . . . 133

6-8 Sample reconstructed images from CMU Multi-PIE dataset using AutoGen. . . . . 134

6-9 Receiver Operating Characteristic (ROC) curves for $24 \times 24$ and $16 \times 16$ resolution face images for the PolyU NIR dataset. Here, 'Proposed' refers to the AutoGen model. . . . . 137

6-10 Sample images from the FG-NET Aging dataset [111]. (a) shows images of individuals below the age of majority, and (b) shows sample individuals at the age of majority. These examples illustrate the challenging nature of *adulthood classification*. . . . . 138

6-11 Sample images from the datasets used for experimental evaluation. . . . . 140

6-12 Receiver Operating Characteristic (ROC) curves obtained for categorizing whether a given face image is of an adult or not. Here, 'Proposed' refers to the Class Specific Mean Autoencoder. Figure has been taken from the published manuscript [141]. . . . . 143

6-13 Sample images from the Multi-Ethnicity dataset, incorrectly classified by all algorithms. At the time of capture, all individuals were below the age of 18. It can be seen that while the actual age of the samples was below the age of majority, it is easy to mistake minors of 16-17 years as adults. External accessories such as scarves may also introduce mis-classification, resulting in unauthorized access control. . . . . 143

6-14 Sample images from the Multi-Ethnicity dataset that are in the age bracket of 16-19 years and misclassified by the proposed Class Specific Mean Autoencoder. The first image belongs to an adult of age 19 years, while the remaining belong to entities below the age of majority. . . . . 144

6-15 Percentage of minors incorrectly classified as adults for both the datasets. A lower percentage would ensure fewer instances of unauthorized access. Here, 'Proposed' refers to the Class Specific Mean Autoencoder. . . . . 145

6-16 Sample images from the MORPH Album-II dataset correctly classified by the proposed algorithm, and not by other existing algorithms. (a) Images of individuals having age 16 (first two samples) or 17. (b) Images of just turned adults of 18 (first two) or 19 years of age. . . . . 146

7-1 This dissertation presents algorithms for facial analysis under low resolution and disguise variations. Potential future research directions have been demonstrated in this figure including (i) utilizing temporal information for rich features via video acquisition, (ii) creating robust automated models invariant to adversarial attacks, (iii) incorporation of ancillary information for improved feature generation, and (iv) development of universal facial analysis models. . . . . 151

THIS PAGE INTENTIONALLY LEFT BLANK

# List of Tables


2.1 Top-1 and top-5 accuracy (%) on the SVHN dataset [104] for VLR digit recognition ($8 \times 8$). . . . . 29

2.2 Rank-1 accuracy (%) for VLR recognition on the CMU Multi-PIE dataset [44]. The HR images are of $96 \times 96$ resolution. . . . . 31

2.3 Rank-1 accuracy (%) on the UCCS dataset [133] for VLR face recognition ($16 \times 16$). The HR images are of $80 \times 80$ resolution. . . . . 31

3.1 Rank-1 identification accuracy (%) on the SCface dataset [42] for LR face recognition. The table presents the performance of the DeriveNet model and the comparative accuracies on the protocol reported by Li *et al.* [81]. . . . . 47

3.2 Rank-1 accuracy (%) on the UCCS dataset [133] for VLR face recognition ($16 \times 16$), with $80 \times 80$ HR images. . . . . 51

3.3 Top-1 and top-5 accuracy (%) on the SVHN dataset [104] for VLR digit recognition ($8 \times 8$). . . . . 52

3.4 Performance of the DeriveNet model with and without the Multi-resolution Pyramid based data augmentation during training. The tabulated metric (%) varies across datasets and is consistent with the ones reported earlier in the manuscript: SCface (avg. rank-1 accuracy), UCCS (rank-1 accuracy), SVHN (rank-1 accuracy), DroneSURF (rank-5 accuracy), and QMUL-SurvFace (avg. accuracy). . . . 54

4.1 Summary of disguise face datasets in literature. . . . . 60

4.2 Statistics of the DFW dataset. . . . . 64

4.3 Statistics of the training and testing sets of the DFW dataset. . . . . 64

4.4 List of teams which participated in the DFW competition. . . . . 68

4.5 Verification accuracy (%) of the participants and baseline performance on protocol-1 (impersonation). . . . . 73

4.6 Verification accuracy (%) of the participants and baseline performance on protocol-2 (obfuscation). . . . . 74

4.7 Verification accuracy (%) of the participants and baseline performance on protocol-3 (overall). . . . . 75

4.8 Number of *easy*, *medium*, and *hard* pairs for 1% and 0.1% FAR. TP and TN refer to True Positive (Genuine) and True Negative (Imposter), respectively. . . . . 77

4.9 Statistics of the DFW2019 dataset. . . . . 80

4.10 Baseline results on the DFW2019 dataset. GAR is reported for the specified FAR values. . . . . 82

4.11 List of teams who participated in the DFW2019 competition. . . . . 84

4.12 Verification accuracy (%) on the proposed DFW2019 dataset for the Impersonation protocol (Protocol-1). The table presents the performance of participants and the baseline results. . . . . 86

4.13 Verification accuracy (%) on the proposed DFW2019 dataset for the Obfuscation protocol (Protocol-2). The table summarizes the performance of participants and the baseline results. . . . . 88

4.14 Verification accuracy (%) for the Plastic Surgery protocol (Protocol-3). Results of the submissions and baseline performance computed using ResNet-50 and LightCNN-29v2 have been presented in the table. . . . . 88

4.15 Verification accuracy (%) for the Overall protocol (Protocol-4). The table presents the performance of the participants and baseline results computed using ResNet-50 and LightCNN-29v2. . . . . 90

4.16 Total *easy*, *medium*, and *hard* pairs at 0.01% FAR. *Easy* refers to the number of pairs correctly classified as TP (True Positive)/TN (True Negative). *Medium* refers to the number of pairs correctly classified as TP/TN by two algorithms, while *Hard* refers to the number of TP/TN pairs correctly classified by at most one algorithm. . 91

5.1 Verification Accuracy (%) on the DFW2018 dataset. Owing to the same protocol, some results have directly been taken from the published paper [143]. The best performance is given in bold, while the second best has been underlined. . . . . 105

5.2 Genuine Acceptance Rate (GAR) (%) on the DFW2019 dataset for Protocol-1 (Impersonation), Protocol-2 (Obfuscation), Protocol-3 (Plastic Surgery), and Protocol-4 (Overall) for two False Acceptance Rates: 0.1% and 0.01%. Comparative results have directly been taken from the published manuscript [136]. The best performance is given in bold, while the second best has been underlined. . . . . 107

5.3 Ablation study of the D-Res framework on the overall protocol of the DFW2018 (Protocol-3) and DFW2019 (Protocol-4) datasets. GAR at 0.1% FAR has been reported, and the efficacy of each component in the framework is evaluated by computing the performance of the framework without the component. . . . . 109

5.4 Verification accuracy (%) of the DED-Net model and comparative techniques on the LFW, YTF, and IJB-B datasets. Comparative results have directly been taken from the different published manuscripts. . . . . 113

6.1 Brief literature review of autoencoder based formulations. . . . . 120

6.2 Classification accuracies (%) for gender classification on $24 \times 24$ and $16 \times 16$ face images from the CMU Multi-PIE dataset. . . . . 130

6.3 Class specific classification accuracies (%) obtained with AutoGen for gender classification on the CMU Multi-PIE dataset. . . . . 131

6.4 Classification performance (%) for gender classification on $24 \times 24$ face images, for the SCface dataset. . . . . 132

6.5 Classification accuracies (%) for gender classification on $16 \times 16$ and $24 \times 24$ face images from the PolyU NIR dataset. . . . . 135

6.6 Classification accuracies (%) for gender classification on $24 \times 24$ NIR face images, for SCface dataset. . . . . 136

6.7 Class specific classification accuracies (%) obtained with **AutoGen** for gender classification on the PolyU NIR dataset. . . . . 136

6.8 Summarizing the dataset description and experimental protocol. . . . . 140

6.9 Classification Accuracy (%) on Multi-Ethnicity dataset. $p$-Value corresponds to the values obtained after performing McNemar test to compare the classification performance of an existing architecture with the proposed Class Specific Mean Autoencoder. The proposed model presents improved classification performance, while being statistically different from all other models at a confidence level of 95%. 144

6.10 Classification Accuracy (%) on MORPH Album-II dataset. $p$-Value corresponds to the values obtained after performing McNemar test to compare the classification performance of an existing architecture with the proposed Class Specific Mean Autoencoder. The proposed model presents improved classification performance, while being statistically different from all other models at a confidence level of 95%. 144

6.11 Confusion matrix of Class Specific Mean Autoencoder on the Multi-Ethnicity database. . . . . 145

6.12 Classification accuracy (%) on perturbed face images for Multi-Ethnicity and MORPH Album-II datasets. . . . . 146

# Acronyms

**CNN** Convolutional Neural Network.

**COTS** Commerical Off-The-Shelf.

**DAE** Denoising Autoencoder.

**DBM** Deep Boltzmann Machine.

**DFW** Disguised Faces in the Wild.

**FPR** False Positive Rate.

**HOG** Histogram of Oriented Gradients.

**HR** High Resolution.

**LBP** Local Binary Patterns.

**LFW** Labeled Faces in the Wild.

**LR** Low Resolution.

**RBM** Restricted Boltzmann Machines.

**RDF** Random Decision Forest.

**ROC** Receiver Operating Characteristic.

**ROI** Region of Interest.

**SVM** Support Vector Machine.

**TPR** True Positive Rate.

**UCCS** UnConstrained College Students (UCCS).

**VLR** Very Low Resolution.

# Research Dissemination

## Publications Related to the Dissertation

### Journals and Selected Conferences:

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

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

3. **M. Singh**, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa, Recognizing Disguised Faces in the Wild, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 2, pp. 97-108, 2019

4. **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*, 2019, pp.340–349

5. **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: 3.255**)

6. A. Sethi, **M. Singh**, S. Nagpal, R. Singh, and M. Vatsa, Residual Codean Autoencoder for Facial Attribute Analysis, *Pattern Recognition Letters*, vol.119, pp.157-165, 2019 (**Impact Factor: 3.255**)

**Other Peer-Reviewed Conference Articles:**

1. **M. Singh**, S. Nagpal, M. Vatsa, and R. Singh, Enhancing Fine-Grained Classification for Low Resolution Images, *International Joint Conference on Neural Networks*, 2021, doi: 10.1109/IJCNN52387.2021.9534026

2. S. Gupta, N. Gupta, S. Ghosh, **M. Singh**, S. Nagpal, M. Vatsa, and R. Singh, A Benchmark Video Dataset for Face Detection and Recognition Across Spectra and Resolutions, in *IEEE International Conference on Automatic Face and Gesture Recognition*, Lille (France), 2019

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

4. **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*, Anchorage (USA), 2017, pp. 1026-1033

5. **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*, Niagara Falls (USA), 2016

**Workshops:**

1. A. Kar, **M. Singh**, M. Vatsa, and R. Singh, Disguised Face Verification using Inverse Disguise Quality, *European Conference on Computer Vision Workshops*, 2020

2. **M. Singh**, M. Chawla, R. Singh, M. Vatsa, and R. Chellappa, Disguised Faces in the Wild 2019, *IEEE International Conference on Computer Vision Workshop on Disguised Faces in the Wild*, Seoul (Korea), 2019

3. **M. Singh**, S. Nagpal, R. Singh, M. Vatsa, and A. Majumdar, MagnifyMe! Aiding Cross Resolution Face Recognition via Identity-Aware Synthesis, *IEEE CVPR Workshop on Biometrics*, Salt Lake City (USA), 2018

4. V. Kushwaha, **M. Singh**, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa, Disguised Faces in the Wild, *IEEE CVPR Workshop on Disguised Faces in the Wild*, Salt Lake City (USA), 2018

## Other Publications

**Journals and Selected Conferences:**

1. S. Nagpal, **M. Singh**, R. Singh, M. Vatsa, N. Ratha, In-group Bias in Deep Learning based Face Recognition Models due to Ethnicity and Age, *Under Submission at IEEE Transactions on Artificial Intelligence*, 2020

2. R. Singh, A. Agarwal, **M. Singh**, S. Nagpal, and M. Vatsa, On the Robustness of Face Recognition Algorithms Against Attacks and Bias, *AAAI Conference on Artificial Intelligence*, 2020

3. S. Nagpal, **M. Singh**, R. Singh, and M. Vatsa, Discriminative Shared Transform Learning for Cross-Domain Matching, *Pattern Recognition*, 2020 (**Impact Factor: 7.196**)

4. **M. Singh**, R. Singh, and A. Ross, A Comprehensive Overview of Biometric Fusion, *Information Fusion*, vol. 52, pp. 187-205, 2019 (**Impact Factor: 13.20**)

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*, Venice (Italy), 2017

**Other Peer-Reviewed Conference Articles:**

1. **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*, 2021 (Accepted)

2. S. Nagpal, **M. Singh**, R. Singh, and M. Vatsa, De-biasing Existing Classification Models Using Diversity Blocks, *IEEE International Joint Conference on Biometrics*, 2020

3. M. Agarwal, S. Sinha, **M. Singh**, S. Nagpal, R. Singh, and M. Vatsa, Triplet Transform Learning For Automated Primate Face Recognition, in *IEEE International Conference on Image Processing*, Taipei (Taiwan), 2019

4. **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*, USA, 2018

5. **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 International Joint Conference on Biometrics*, Denver (USA), 2017(**Received Best Poster Award**)

6. S.Nagpal, **M. Singh**, R. Singh, M. Vatsa, and A. Noore, On Matching Skull to Digital Face Images: A Preliminary Approach, *IEEE International Joint Conference on Biometrics*, Denver (USA), 2017, pp. 813-819

7. 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*, Anchorage (USA), 2017, pp. 3814-3821

8. S. Yadav, **M. Singh**, M. Vatsa, R. Singh, and A. Majumdar, Low rank group sparse representation based classifier for pose variation, *IEEE International Conference on Image Processing*, Phoenix (USA), 2016, pp. 2986-2990

**Workshops and Book Chapters:**

1. 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*, 2020

2. S. Nagpal, **M. Singh**, M. Vatsa, R. Singh, and A. Noore, Expression Classification in Children Using Mean Supervised Deep Boltzmann Machine, *IEEE CVPR Workshop on Analysis and Modeling of Faces and Gestures*, Long Beach (USA), 2019

3. T. Chugh, **M. Singh**, S. Nagpal, R. Singh, and M. Vatsa, Transfer Learning based Evolutionary Algorithm for Composite Face Sketch Recognition, *IEEE CVPR Workshops*, Hawaii (USA), 2017, pp. 619-627

4. S. Nagpal, **M. Singh**, M. Vatsa, and R. Singh, Deep Learning: Fundamentals and Beyond, *Deep Learning for Biometrics*, CRC Press, 2017

# Chapter 1

# Introduction

Facial analysis often comes naturally to humans, wherein we are able to estimate several attributes from the facial region with great precision and ease [146]. As shown in Figure 1-1, attributes such as the gender, age, or race are often estimated quickly by us, along with the identity information, if previously known. While the knowledge of such attributes facilitates smoother inter-personal interactions in day-to-day setups, they are also imperative for person description and identification, enhanced customer service in business setups, and for person authentication in law enforcement scenarios. Owing to the utility of such attributes and identity information in different situations, research in the past few decades has focused extensively on automating the said task of facial analysis. Automated facial analysis has wide-spread applicability including authentication for authorized access, attendance monitoring, expression classification for distress identification or attention monitoring, and even disease diagnosis.

Figure 1-2 presents a traditional facial analysis pipeline involving multiple stages such as: (i) data acquisition, (ii) facial detection, (iii) pre-processing such as geometric alignment, (iv) feature extraction such as hand-crafted (Histogram of Gradients [19] or Local Binary Patterns [109]) or deep-learning based, and (v) classification via models such as a Support Vector Machine or a Neural Network. Initial research focused heavily on analyzing face images captured in *constrained settings* with good illumination, frontal pose, and neutral expressions (e.g., CMU Multi-PIE [44], Yale [8], and AR [91] datasets). With significant progress in the area of Computer Vision and Machine Learning, facial analysis algorithms began achieving high performance on such constrained images and datasets useful for authentication/analysis in controlled setups.

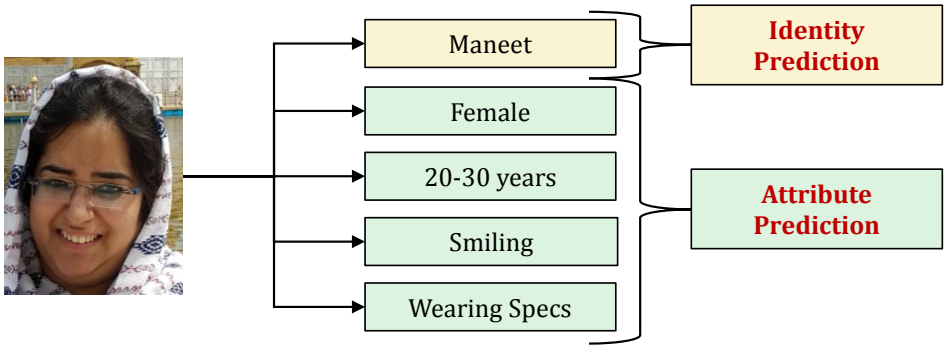



Figure 1-1: Facial analysis refers to the task of predicting the identity or attribute information (such as age, expression, or gender) from a given face image.

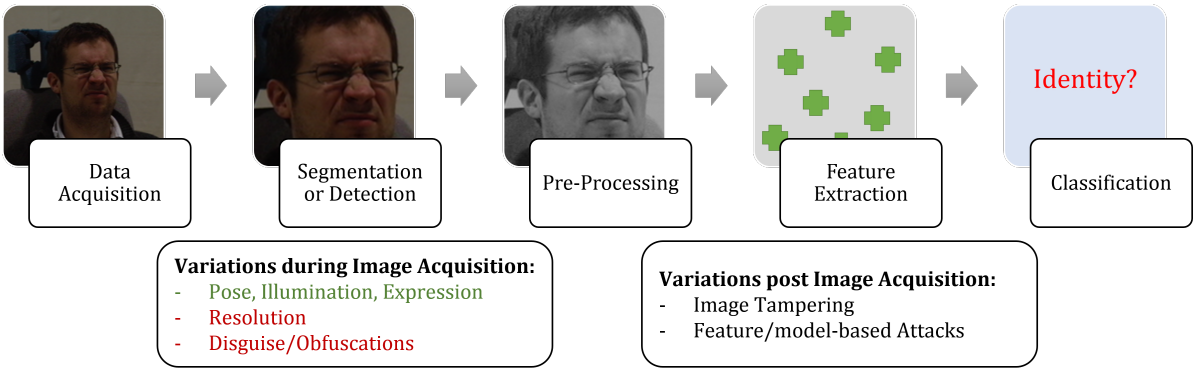



Figure 1-2: Sample traditional facial analysis pipeline consisting of different stages. In recent algorithms (such as deep-learning based techniques) some of the steps are combined into one (e.g., a Convolutional Neural Network is used for feature extraction and classification). Often, variations are encountered by a facial analysis pipeline either during image image acquisition (e.g., changes due to pose, resolution, or usage of disguise accessories) or post the image acquisition step (e.g., image/feature tampering or model-based attacks).

Beyond controlled setups, automated facial analysis models used in real-world scenarios are often susceptible to variations in the input image either before or post the image acquisition (Figure 1-2). Variations prior to the image acquisition alter the captured image resulting in a relatively lesser constrained facial region, while variations post image acquisition often tend to target the image processing/classification pipeline for altered predictions. Post success in constrained facial analysis, research efforts were directed towards identifying different *covariates* and analyzing facial images captured in relatively *unconstrained settings* with variations to the facial region. Figure 1-3 presents an overview of an authentication system in real-world scenarios. Often, the model is trained with images captured in constrained settings such as high illumination, neutral expression,


Figure 1-3: Often in real-world scenarios, the authentication system is developed using facial images captured in constrained environments. However, at real-time the system might encounter images captured in unconstrained scenarios with several covariates such as resolution variations or presence of disguise accessories. The authentication system must therefore be able to match the given input with the constrained gallery samples provided during enrollment.

and frontal pose. Post deployment, the trained model is often expected to correct identify (or analyze) facial images with varying covariates as well. In order to perform well, the algorithm must therefore be able to extract features invariant to the variations observed due to the different covariates. To this end, research efforts focused on addressing such requirements by developing algorithms capable of facial analysis on images with variations in pose, illumination, and expression. This was followed by identification of several covariates which affect facial analysis algorithms, such as the effect of **(i)** age, **(ii)** disguise, **(iii)** make-up, **(iv)** resolution, and **(v)** plastic surgery (e.g., FG-Net [111], Plastic Surgery [144], and IIITD $I^2$BVSD [24] datasets). Parallely, research also focused on *cross-domain identification algorithms* for scenarios such as sketch to digital image matching, high resolution to low resolution matching, and visible to NIR spectrum matching. However, most of the research still focused on addressing the effect of a single covariate for facial analysis. One of the seminal works in the progress of facial analysis was the introduction of the Labeled Faces in the Wild (LFW) dataset [57] which provided *unconstrained* images captured from the Internet. The dataset was able to capture simultaneous variations across pose, illumination, expressions, and make-up, captured from different sources. High performance was soon obtained on the dataset, followed by the exploration of combination of other covariates for analysis.

Despite the several advancements, while humans achieve superlative performance across such different real-world scenarios on a daily basis, automated facial analysis is yet to attain similar heights. For instance, while near perfect performance has been obtained on the LFW dataset, the performance on the challenging real-world QMUL SurvFace [18] dataset or the SCFace dataset

(a) (b)

(c)

Figure 1-4: Images obtained from the surveillance feed of CCTV cameras related to two unfortunate incidents: (a) Frame captured on a New Delhi metro station depicting two people after abandoning a newborn girl child at the metro station [98], (b) Suspect of the infamous Chennai murder case, where a young woman was murdered at the local railway station [107]. (c) Sample images of the suspect involved in the Boston Marathon Bombing, 2013. All the images suffer from the challenge of low resolution region of interest (facial region), while (c) also suffers from the presence of a disguise accessory (cap).

[42] remains to be below 80% [95]. Datasets such as these replicate the real-world scenario of face images captured from a distance, often via a CCTV or surveillance camera. Facial analysis in such scenarios is of utmost importance since it can help facilitate law-enforcement agencies during crisis by assisting in attribute and identity prediction of face images captured from a distance. For example, Figure 1-4 presents sample images of recent incidents caught on camera, where the person of interest is far away from the camera, and is seen to often utilize a disguise accessory. Specifically, Figure 1-4(c) refers to the unfortunate incident of Boston Marathon Bombing (2013), where cameras were able to capture images of the suspect, however, it was the manual intervention and search that led to the identification of the suspect. Automated facial analysis can thus assist in identification of the individual or raise an alarm for sightings at other public locations. Data

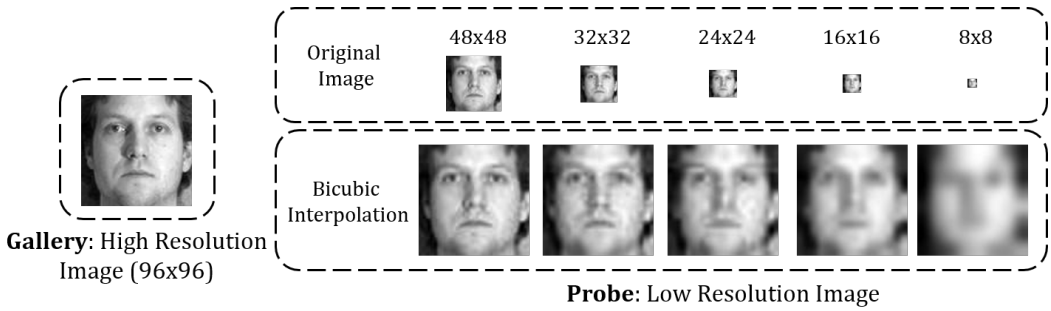



Figure 1-5: Variations observed in facial region at different resolutions: $48 \times 48$, $32 \times 32$, $24 \times 24$, $16 \times 16$, and $8 \times 8$. Reduced resolution results in lesser fine details, low quality, and limited interpretative information content.

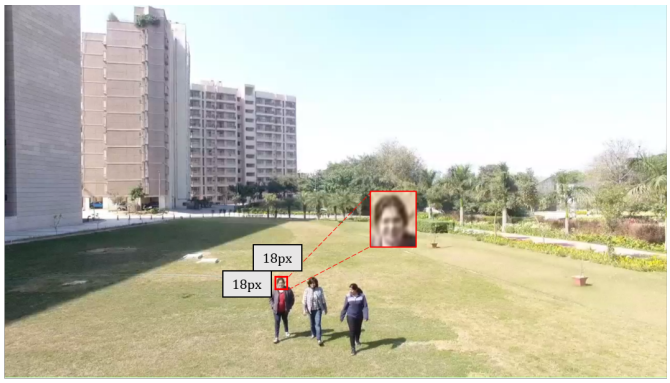
captured in such scenarios typically suffer from two key challenges (which have been relatively less explored in literature) along with other covariates:

- *low or very low resolution* of the facial region, and
- utility of *disguise accessories* by individuals.

Both the factors result in variations to the facial region as compared to the one captured in a constrained high resolution setup, and correspond to variations before the image acquisition (Figure 1-2). Further, facial analysis under low resolution or disguise variations also has applicability in scenarios where multiple people are captured in the same image, resulting in a reduced effective face resolution with application to image tagging on social media or photo management applications. To this end, this dissertation focuses on developing facial analysis solutions for images captured under low resolution or disguise variations. The remainder of this chapter elaborates upon the two challenges in detail, followed by the research contributions of this dissertation.

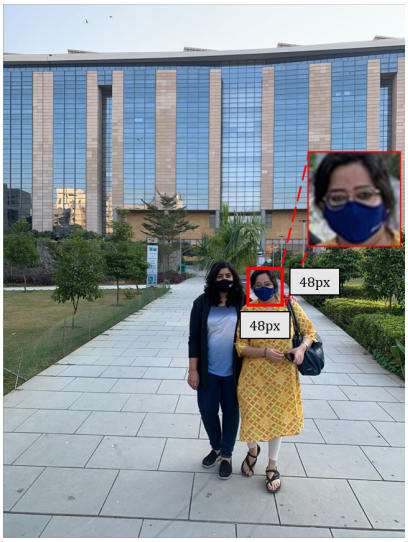
## 1.1 (Very) Low Resolution Facial Analysis

The resolution of an image is often synonymous with the *level of detail* it holds, and is often quantified by the *number of pixels* present in the image. Figure 1-5 presents a sample face image at different resolutions and its corresponding bicubic interpolated image at $96 \times 96$. It can be observed that as the resolution of the image reduces, it captures less details, and thus contains fewer discriminative biometric identifiers of reasonable clarity, making it challenging for auto-




(a) Frame: 1280x720


(b) Image: 1600x1000




(c) Image: 850x640

Figure 1-6: Sample images demonstrating low resolution regions of interest for facial regions.

mated analysis. Generally, in the literature, low resolution (LR) classification refers to identifying image regions with a resolution of $32 \times 32$ to $16 \times 16$ pixels. When the resolution of the region of interest is even lower than $16 \times 16$ pixels, the task is referred to as very low resolution (VLR) recognition/classification [201].

Most of the existing deep learning based models utilize input images with facial regions of dimension around $224 \times 224$, which is often not available in real world scenarios. While existing hand-held devices (cameras or smartphones) are being able to capture images of at least $2560 \times 1960$ resolution (5MP), the effective region of interest (ROI) in the captured image could be much lower. For example, Figure 1-6 presents sample scenarios wherein the effective resolution of face images is often less than $64 \times 64$, despite the higher resolution of the complete image:

- Figure 1-6(a) presents a sample frame captured via a drone flying at a height. Drone-based facial analysis has recently garnered research focus [64], owing to its applicability in several

critical scenarios such as supporting law-enforcement in large-scale events, person identification/detection in challenging terrains, and missing person identification. Recently, in 2019, IARPA also began the Biometric Recognition and Identification at Altitude and Range (BRIAR)[1] program for encouraging research on biometric recognition from high altitude and range.

- Figure 1-6(b) presents a group photograph captured from a high-end smart-phone, wherein the effective facial region is less than $32 \times 32$.

- Figure 1-6(c) presents a sample image captured from a hand-held device, where the subject is present at a distance of almost $10m$ from the device. Both the above scenarios are often frequently encountered in our day-to-day lives, and the automated processing of such facial regions also requires modeling the observed resolution variations.

Beyond these scenarios, low resolution or very low resolution images can also be a result of the inherent limited resolution of the acquisition device. For example, CCTV cameras used for surveillance often do not capture very high resolution images, which combined with the large field of view and distance of the subject, renders a very low resolution ROI. Due to the easy availability and installation, CCTV cameras are now being installed in several public areas such as parks, markets, and community centers for providing enhanced security. As of 2019, it is estimated that the United States of America has 15.28 CCTV cameras for every 100 individuals, while the United Kingdom has seven [23]. Such devices capture huge amount of data, which can effectively be utilized in critical scenarios of missing person identification or suspect tracing.

While facial analysis on LR or VLR facial regions hold substantial importance, it is yet to achieve high performance as observed in constrained HR facial analysis, or even analysis under the effect of single covariates (such as pose, expression, or illumination). As observed from Figure 1-5, there is a vast difference in the appearance of images at varying resolutions. Thus, a facial analysis pipeline (Figure 1-2) developed only for HR samples is often unable to extract meaningful features for LR samples. Traditional facial analysis models relied on extracting fixed-length handcrafted features (e.g., HOG, LBP) for the given input sample (of a fixed resolution). Recently,

[1] https://www.iarpa.gov/index.php/working-with-iarpa/requests-for-information/biometric-recognition-and-identification-at-altitude-and-range-briar

with deep-learning algorithms achieving very high performance, most of the models are trained for high resolutions samples (often having resolution greater than $224 \times 224$). In order to utilize such models/algorithms, the LR or VLR samples have to be super-resolved to the input dimension, which often results in the loss of key biometric features useful for recognition. *Large magnification factors* (e.g., a magnification factor of 16 is required for a $14 \times 14$ VLR input image) can especially result in the introduction of noise and other artifacts in the super-resolved images. Therefore, direct utilization of HR models for VLR/LR face images often results in undesirable performance [142]. Depending upon the acquisition device and stand-off distance, VLR/LR samples also contain limited interpretive information content as compared to HR samples, specifically with respect to detailed biometric components and features. Further, the lower resolution of images is often accompanied with other degradation factors due to blurring, sensor noise, or compression artifacts introduced during image acquisition or storage.

The above mentioned challenges render the task of LR/VLR image classification further arduous. In order to address the given problem, analysis models can be developed utilizing VLR/LR samples only. Such models are often challenging to learn since LR/VLR images contain *limited interpretive information* resulting in ineffective feature extraction. Further, the inter-class and intra-class variations are also not as pronounced in VLR/LR samples as compared to their corresponding HR samples. Owing to these limitations and also the availability of HR samples, most of the existing techniques in literature utilize HR samples during the training of LR/VLR analysis models. While utilizing both HR and VLR/LR samples, three types of algorithms have been proposed:

- **Transformation based approaches:** In transformation based approaches, a transformation function is learned between the HR space and the VLR/LR space. The transformation can be applied either at the image-level (super-resolution or down-sampling techniques) or at the feature-level.

- **Resolution-invariant model learning:** Resolution-invariant model learning focuses on learning an analysis pipeline which is able to extract resolution-invariant features and learn a classification model invariant to the input resolution.

- **Learning from auxiliary HR samples for better classification:** The third category of tech-

Figure 1-7: Utilization of disguise accessories which obfuscate the facial region often result in large intra-class variations.

niques focus on adapting the high structural and detailed information available in some auxiliary HR images for learning an effective LR/VLR classifier.

Despite the increasing research attention on VLR/LR analysis tasks, the state-of-the-art models still suffer from low classification performance. The arduous nature of the problem, along with its wide-scale applicability in real-world scenarios demands further dedicated research.

## 1.2 Disguised Face Recognition

The second challenging yet less explored covariate in unconstrained facial analysis is authentication under the presence of *disguise variations*. As can be observed from Figure 1-7, images of an individual vary in appearance based on the type of disguise accessory used. As is the case with other covariates (Figure 1-3), an automated facial recognition model is often faced with the challenge of matching a gallery image captured in constrained settings (e.g., first image in Figure 1-7) with disguised images obtained during real-time (images 2-6 of Figure 1-7). The variations observed in the facial images with and without a disguise accessory make it challenging for an automated recognition system to match accurately.

The task of disguised face recognition is further characterised by the availability of different disguise techniques. One of the most common ways of disguising the facial region is via the use of external artifacts (such as hat, cap, scarf, sunglasses, and muffler). Beyond such artifacts, the facial region can also be disguised by means of make-up, plastic surgery, and even natural alterations such as hair changes (growth or removal of beard/moustache), sudden weight gain/loss, etc. Further, the usage of such techniques is also governed by the notion of *intent*:

- Changes in the facial region might result due to an unintentional utility of an accessory.

(a) Accessories resulting in Obfuscation


(i) Person-A, (ii) Passport of Person-B, (iii) A *impersonating* B

(b) Recent Incident of Impersonation

Figure 1-8: Day-to-day accessories can often result in (a) obscuring different facial regions [136], while such accessories can also be used to (b) impersonate another individual [108], resulting in challenges for an automated face recognition system.

For example, sunglasses during Summers, mufflers during Winters, or face masks during pollution (Figure 1-8(a)).

- Beyond unintentional usage, disguise can also be used for obtaining unauthorized access by intentionally impersonating another individual. Figure 1-8(b) presents a recent incident where Person-A impersonated Person-B for obtaining unauthorized access at an airport. The impersonation attempt was caught due to the manual intervention and verification of the physical identification document.

Owing to the large and (often) unintentional use of such accessories, images captured in unconstrained settings (such as from a distance or via surveillance feed) often contain disguise accessories, thus requiring research focus for accurate recognition.

Regardless of the disguise technique, most of the disguises result in the *obfuscation* of the face. Further, each technique (such as accessories such as hats, caps, scarves, sunglasses, masks, and moustache) result in the obfuscation of different facial regions with varying extent. Traditional facial analysis algorithms utilize the entire face for feature extraction and classification, and thus

often fail with the recognition of disguised faces due to the non-biometric information present in the region of interest. The problem is further exacerbated for images captured from a distance having LR/VLR region of interest (facial region). For example, Figure 1-6(c) presents a sample image captured from a distance, where the subject has also worn a face mask, thus obfuscating the facial region. In such real-world scenarios (also discussed in the previous Section), the captured images also suffer from resolution variations. Coupled with enhanced intra-class variations (due to obfuscation) and reduced inter-class separability (due to impersonation), the task of disguised face recognition is rendered further challenging.

Research in the area of disguise face recognition began with datasets captured in constrained settings (neutral expression, well illumination, frontal pose, and high resolution) having a single disguise accessory. For example, early literature utilized the AR dataset [91] (containing sunglasses and scarf) and the Yale face dataset [8] (containing only sunglasses). Initial algorithms on disguised face recognition focused mostly on matching features extracted from the biometric (non-disguised) regions of the faces only [92, 118]. Due to the relatively constrained datasets, research soon witnessed algorithms achieving very high performance for disguised face recognition. In 2013, Dhamecha *et al.* [24] released a novel IIITD $I^2$BVSD dataset which covered a wide range of disguise accessories such as hats, wigs, scarves, sunglasses, masks, etc. The dataset captured face images in the visible and thermal spectrum, in a well illuminated controlled environment, with frontal pose and neutral expression. This was followed by the release of more datasets such as the Disguise and Makeup dataset [167] and the Spectral Disguise Face Dataset [119], both of which contained facial images with varying disguise accessories. The former was collected from the Internet, while the later was captured in controlled laboratory settings. Recently, with the advent of representation learning algorithms, researchers have also tried to learn disguise-invariant representations for efficient face recognition. Moreover, the availability of real-world datasets and challenging competitions, such as the Disguised Faces in the Wild (DFW) competition series [136, 143], have further pushed the state-of-the-art for disguised face recognition.

Despite the recent advances and improvement in disguised face verification, the current state-of-the-art performance is substantially lower as compared to traditional face recognition. For example, the best reported results for the *Impersonation* protocol of the DFW 2019 dataset are around 78% [136], whereas over 99% verification performance has already been achieved on the widely-

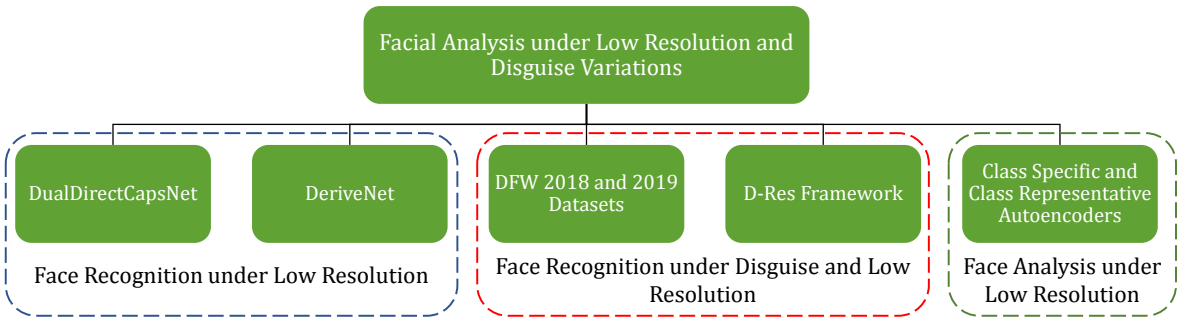



Figure 1-9: Key contributions of this thesis focusing on facial analysis under low resolution and disguise variations.

used Labeled Faces in the Wild (LFW) dataset [57]. Further, to the best of our knowledge, no research has focused on understanding the effect of resolution on disguised face verification; an imperative problem for unconstrained images captured from a distance (or in surveillance settings).

## 1.3 Research Contributions

As elaborated in the previous Sections, two key covariates for unconstrained facial analysis have received limited research attention: (i) resolution variations (resulting in VLR/LR facial regions) and (ii) disguise variations. To this effect, this dissertation focuses on developing algorithms for facial analysis under low resolution and disguise variations (Figure 1-9). Specifically, deep learning algorithms have been presented for low and very low resolution facial recognition and analysis, along with a face verification framework for recognizing faces under disguise variations. Two novel unconstrained disguised face datasets have been prepared and released which help in extending the current body of literature by facilitating research in the challenging problem of disguise face verification. Further, this dissertation also focuses on novel tasks such as attribute prediction and disguise face recognition in low resolution setups. Deep learning based algorithms are presented which enable learning rich representations capable of modeling the challenges observed across the covariates. The presented algorithms have been evaluated on standard publicly available benchmark datasets, and comparisons have been made with the state-of-the-art techniques. The presented algorithms demonstrate improvement as compared to the state-of-the-art performance on various datasets. The key contributions of this dissertation are:

- **Dual Directed Capsule Network for VLR Face Recognition:** VLR face recognition corresponds to identifying facial regions with resolution $16 \times 16$ or less. This dissertation presents a novel *Dual Directed Capsule Network* model, termed as *DirectCapsNet*, for addressing VLR face recognition. The proposed architecture utilizes a combination of capsule and convolutional layers for learning an effective VLR recognition model. The architecture also incorporates two novel loss functions: (i) the HR-anchor loss and (ii) the targeted reconstruction loss, which enable the model to learn from some auxiliary HR images available during training. Multiple experiments for VLR face recognition are performed along with comparisons with state-of-the-art algorithms. The DirectCapsNet model demonstrates improved recognition performance as compared to the state-of-the-art results on different benchmark datasets.

- **DeriveNet for VLR/LR Face Recognition:** As mentioned previously, VLR/LR images (or regions of interest) often contain less information content, rendering ineffective feature extraction and classification. This dissertation presents a novel *DeriveNet* model for VLR/LR classification, which focuses on learning effective class boundaries by utilizing the class-specific domain knowledge. The DeriveNet model is jointly trained via two novel losses: (i) Derived-Margin softmax loss and (ii) the Reconstruction-Center (ReCent) loss. The Derived-Margin softmax loss focuses on learning an effective VLR classifier while explicitly modeling the inter-class variations. The ReCent loss incorporates domain information by learning a HR reconstruction space for approximating the class variations for the VLR/LR samples. It is utilized to *derive* inter-class margins for the Derived-Margin softmax loss. Experiments and analysis have been performed on multiple datasets, including the DroneSURF dataset for *VLR/LR face recognition from drone-shot videos*. The DeriveNet model achieves state-of-the-art performance across different datasets, thus promoting its utility for several VLR/LR classification tasks.

- **Disguised Faces in the Wild Benchmark Datasets:** Research in the area of disguised face recognition has been restricted by the presence of limited unconstrained in-the-wild datasets. In order to facilitate further research, as part of this dissertation, *two challenging datasets* have been presented: (i) the Disguised Faces in the Wild (DFW) 2018 dataset, and (ii) the

DFW 2019 dataset. The DFW2018 dataset was released as part of the First International Workshop and Competition on Disguised Faces in the Wild at the *International Conference on Computer Vision and Pattern Recognition*, 2018, containing over 11,000 images of 1,000 identities with variations across different types of disguise accessories. The DFW2019 dataset was released as part of the Second International Workshop and Competition on Disguised Faces in the Wild at the *International Conference on Computer Vision*, 2019, containing 3840 images of 600 subjects. All images are collected from the Internet via relevant keyword searches on different search engines, thereby demonstrating wide variations with respect to pose, illumination, lighting, resolution, capturing device, and disguise accessories. The DFW2019 dataset contains variations due to different disguise accessories, and before-after images for plastic surgery and bridal make-up. This dissertation presents the two datasets in detail, including the evaluation protocols, baseline results, and the performance analysis of the submissions received as part of the competition.

- **Disguise Resilient Face Verification:** As discussed previously, external artifacts and makeup accessories may result in the obfuscation of one's identity. Such accessories may also intentionally be used to impersonate someone else's identity, often rendering automated facial recognition systems ineffective. To this effect, this dissertation presents a novel *multi-objective encoder-decoder network*, termed as DED-Net, for learning disguise invariant features by means of different distance metrics (Mahalanobis and Cosine distance), along with a Mutual Information based supervision. The DED-Net has been extended to present the Disguise Resilient (D-Res) framework which combines local and global features for accurate face recognition with disguise variations. The efficacy of the framework has been demonstrated on two real-world benchmark datasets: Disguised Faces in the Wild (DFW) 2018 and DFW2019 competition datasets. Additionally, to the best of our knowledge, this is the first research in literature which emphasizes on handling *disguised faces in low resolution settings* to simulate real-world surveillance scenarios. Benchmark results have been shown on seven protocols for three low resolution settings ($32 \times 32$, $24 \times 24$, and $16 \times 16$) of the two DFW benchmark datasets. Superior performance in comparison with benchmark and state-of-art algorithms presents the effectiveness of the presented framework.

- **Attribute Prediction from VLR/LR Facial Regions:** Attribute prediction from facial images has wide-spread utility in scenarios related to human computer interaction, marketing, security, and demographic reporting. Extensive research has focused on predicting attributes (such as gender, age) from HR face images, however, limited attention has been given to developing algorithms for attribute prediction from VLR/LR facial regions. This dissertation presents two deep learning based autoencoder formulations for learning rich representations while explicitly modeling the inter-class and intra-class variations. The presented *Class-Specific Mean Autoencoder* and *Class Representative Autoencoder* demonstrate improved performance as compared to the other comparative techniques for the task of gender and adulthood prediction from facial regions.

THIS PAGE INTENTIONALLY LEFT BLANK

# Chapter 2

# Dual Directed Capsule Network for Very Low Resolution Image Recognition

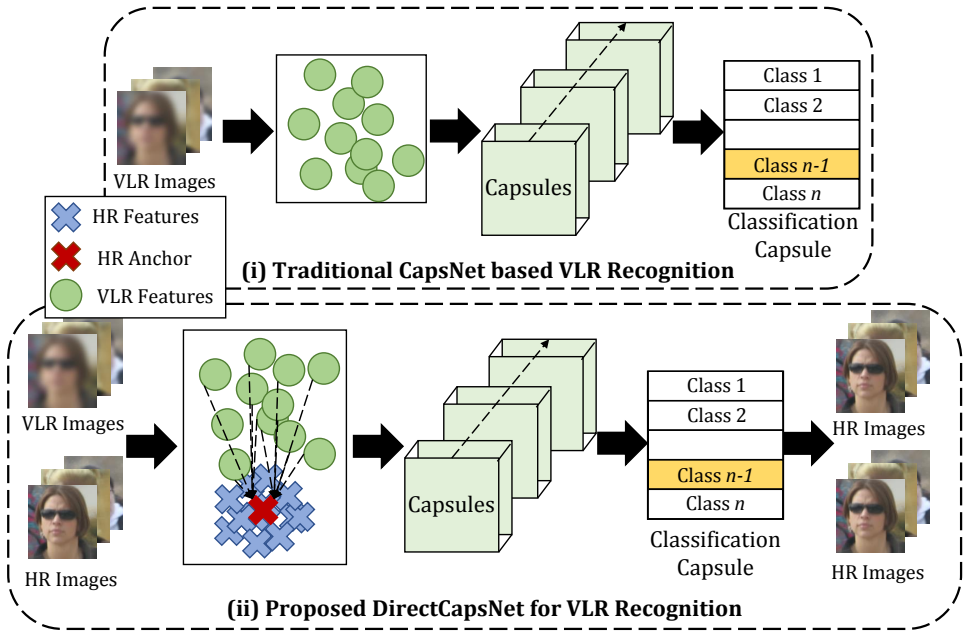
## 2.1 Introduction

In typical surveillance scenarios, images are often captured from a large stand-off distance, thus rendering the region of interest to be of a very low resolution (VLR), often times less than $16 \times 16$ [201]. Figure 2-1(a) shows sample real-world applications of VLR recognition where the region of interest can be a face, a suspicious object, or the license plate number of a moving vehicle. These samples demonstrate the arduous nature of the problem where some of the key challenges of VLR recognition are the presence of limited information content and blur. VLR recognition also has applicability in image tagging, where multiple objects/people are captured in the frame, and each of these entities are of small resolution.

Netzer *et al.* [104] demonstrated the poor performance of humans on identifying VLR digits captured in real surroundings. For the Street View House Numbers (SVHN) dataset, the authors observed cent percent accuracy by humans for samples with $101 - 125$ pixel height. On the other hand, the performance dropped to $82.0\% \pm 2\%$ when classifying very low resolution samples i.e., images of height up to 25 pixels, thereby reinstating the challenging nature of the problem. Direct up-sampling via interpolation could be viewed as a possible solution for VLR recognition, however, multiple studies have demonstrated poor performance owing to the required large magnification factor [81, 142] and possible introduction of noise, which can also be observed in Figure


(i) Digit Classification (ii) Face Recognition

(a) Real-world applications of VLR recognition. Image source: (i) Internet, (ii) UCCS dataset [133]




(b) Proposed Dual Directed Capsule Network (DirectCapsNet)

Figure 2-1: The proposed DirectCapsNet utilizes HR samples to *direct* learning of more meaningful and discriminative features for VLR image recognition via the proposed HR-anchor loss and the targeted reconstruction loss.

2-1(a)(i). Further, in the literature, researchers have also demonstrated the inability of models trained on high resolution (HR) images (containing high information content) to perform well on (V)LR images [142]. The current state of scarce available solutions and the wide applicability of VLR recognition makes it an important problem, demanding dedicated attention.

This research proposes a novel capsule network based model for VLR image recognition. Hinton *et al.* [51] proposed learning "capsules", which represent a vector of instantiation parameters in order to encode the input more efficiently. Instantiation parameters may constitute the properties

of an image such as the *pose, lighting, and deformation of the visual entity relative to an implicitly defined canonical version of that entity* [51]. We believe that such parameters would be invariant to the resolution of the image, therefore presenting the potential of being useful for VLR recognition. Due to the limited information content in VLR images, the VLR recognition model could benefit from the information-rich HR samples as well. To this effect, we propose *Dual Directed Capsule Network* (termed as *DirectCapsNet*) (Figure 2-1(b)) to learn meaningful features for VLR recognition, directed (or guided) by the HR samples. The contributions of this research are as follows:

- A novel Dual Directed Capsule Network (*DirectCapsNet*) model is proposed for VLR recognition, which *directs* the features learned from the VLR images containing limited information towards the more meaningful and discriminative features of the HR images.
- Two losses are proposed for directing the VLR recognition model: (i) HR-anchor loss and (ii) targeted reconstruction loss. HR-anchor loss is proposed for the feature learning module, which pushes the VLR features of a particular class towards a representative HR feature (anchor) of that class. Targeted reconstruction loss is utilized at the classification module, where HR images are reconstructed from the capsule outputs of the VLR images, thereby forcing the capsules of VLR and HR images of the same class to be similar.
- Experimental results and analysis demonstrate the advantages of the proposed DirectCapsNet model for VLR digit classification and VLR face recognition. Experiments are performed on the SVHN [104], CMU Multi-PIE [44], and UCCS [133] databases, and comparisons are performed with state-of-the-art algorithms. The proposed model yields over 95% accuracy on the challenging UCCS face database. On the SVHN database, it achieves about 84% classification accuracy with $8 \times 8$ VLR images demonstrating an improvement of almost 27% from the existing results.

## 2.2 Related Work

There have been several advances in the field of low resolution recognition [61, 81, 110, 171]; however, the area of very low resolution (VLR) recognition remains relatively less explored. As

mentioned previously, very low resolution (VLR) recognition refers to identifying regions of interest with $16 \times 16$ resolution or less. Owing to the limited information content in a given VLR image, a potential solution is to super-resolve or synthesize its higher resolution image [112, 166], which is then used for recognition. While there exists vast literature on super-resolution or synthesis algorithms [78, 132, 169], most of them focus primarily on the visual quality of the generated image, and not on the task of recognition. Zou and Yuen [201] proposed one of the initial super resolution techniques with specific focus on VLR face recognition. The proposed algorithm utilizes a combination of visual quality based constraint for good quality HR synthesis, and a discriminative constraint for learning features useful for recognition. Singh *et al.* [142] proposed an identity-aware face synthesis technique for generating a HR image from a given LR input. The synthesized images were provided to a Commercial-Off-The-Shelf (COTS) system for recognition.

Apart from super-resolution based techniques, in the literature, researchers have also proposed algorithms for *enhancing* or *improving* the features learned for VLR images by using the information extracted from the HR images. For instance, Bhatt *et al.* [12] proposed an ensemble-based co-transfer learning algorithm for face recognition. The co-transfer algorithm operates at the intersection of co-training and transfer learning by utilizing the information of HR images for enhancing the VLR classification. Wang *et al.* [170] proposed Robust Partially Coupled Networks for VLR recognition. HR images are used as "auxiliary" data during training for learning more discriminative information, which might not be available in VLR images. As demonstrated via multiple experiments, using HR images at the time of training, enhances the learned VLR features, resulting in improved recognition performance. Mudunuri and Biswas [102] proposed a reference-based approach along with multidimensional scaling for learning a common space for HR and VLR images. Recently, Li *et al.* [81] analyzed different metric learning techniques for LR and VLR face recognition, by learning a common feature space for HR and LR samples. Ge *et al.* [37] proposed a selective knowledge distillation technique for (V)LR face recognition. A base network trained on HR face images is used for selecting the most informative facial features for a (V)LR CNN model, in order to enhance the (V)LR features and the classification performance.

In the literature, VLR recognition algorithms have shown to benefit from HR samples by learning shared representations between the HR and VLR samples [170] or by transferring the model information learned by the HR data onto the VLR recognition model [37]. By utilizing the addi-

tional information from the HR images at the time of training, such algorithms are able to learn more discriminative and meaningful features, as compared to those learned independently from the VLR images. This research proposes to utilize the auxiliary HR samples during training to *direct* the VLR features towards the more informative HR features, via a novel DirectCapsNet model.

## 2.3 Proposed Dual Directed Capsule Network

As shown in Figure 2-2, the problem of very low resolution (VLR) recognition suffers from the challenge of limited information content in the input images, which often results in lack of discriminative features useful for recognition/classification. In order to overcome this challenge, we propose a novel Dual Directed CapsNet, termed as *DirectCapsNet*. DirectCapsNet enhances the VLR representations by directing them in two ways: via the proposed (i) HR-anchor loss and (ii) targeted reconstruction loss, both of which provide additional supervision using the HR images. The HR information is used to direct/guide the framework to extract discriminative representations even from the VLR images having limited information content. This is accomplished by using the HR-anchor loss which brings the representations of VLR images closer to the representations of their corresponding HR samples. This is also enforced at the classification stage via the targeted reconstruction loss, which promotes similar features for HR and VLR samples of the same class. Since the base architecture of the proposed model is a capsule network, we first briefly explain its functioning, followed by the in-depth explanation of the proposed model.

### 2.3.1 Preliminaries: Capsule Networks

Hinton *et al.* [51] proposed the concept of *capsules* as an effective method of learning representations. It was further developed by Sabour *et al.* [131], where a capsule network (CapsNet) is presented for classification. A capsule is a "group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part". In other words, instead of a single scalar output, each capsule outputs a vector, the values of which are referred to as the activity vector. The length of each capsule vector ($\|.\|_2$) is bounded in the range of $[0-1]$. Sabour *et al.* [131] proposed the concept of dynamic routing between capsules, wherein multiple layers of capsules were stacked for object classification. The final layer contains the clas-

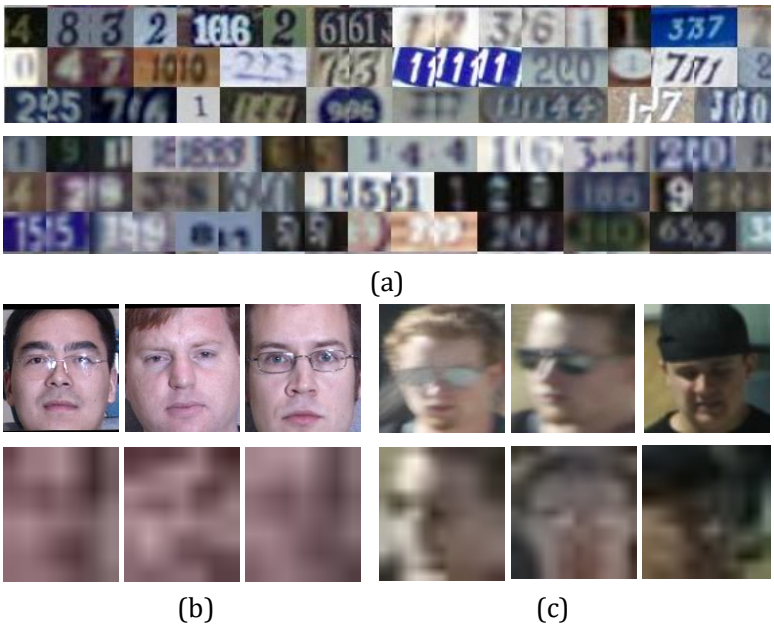



Figure 2-2: Sample HR and VLR images from the (a) SVHN dataset, (b) CMU Multi-PIE dataset, and (c) UCCS dataset. The HR images (first row) contain high information content, which is often missing in the VLR samples (second row).

sification capsules of dimension $k \times m$, where $k$ is the number of classes and $m$ is the capsule dimension. For a given input, the predicted class is the class corresponding to the capsule with the maximum activity vector (length). In order to learn an effective classification model, margin loss is used to learn the network. Given a $K$ class problem, with $v_k^{x^c}$ as the output of the $k^{th}$ class capsule for an input $x^c$ (belonging to class $c$), and $T_k$ being the label corresponding to the $k^{th}$ class, the margin loss of CapsNet is defined as:

$$\mathcal{L}_{Margin} = \sum_{k=1}^{K} \left( T_k \max(0, m^+ - \|v_k^{x^c}\|)^2 \; + \; \lambda(1 \; - \; T_k) \max(0, \|v_k^{x^c}\| - m^-)^2 \right) \tag{2.1}$$

where, $T_k \in \{0, 1\}$, that is, whether the input sample belongs to class $k$ ($T_k = 1$) or not ($T_k = 0$). $m^+$ and $m^-$ correspond to the positive and negative margin used to increase the intra-class similarity and reduce the inter-class similarity, respectively, and $\lambda$ is a constant for controlling the weight of each term. The above loss (Equation 2.1) promotes a larger length of capsule ($\|v_k\|$) for the correct class, and a smaller length for capsules corresponding to the other classes. Capsule networks are relatively less explored in the literature, with limited or no modification to the architecture or loss function. They have been used for brain tumor detection [2], sea grass detection [59], generat-

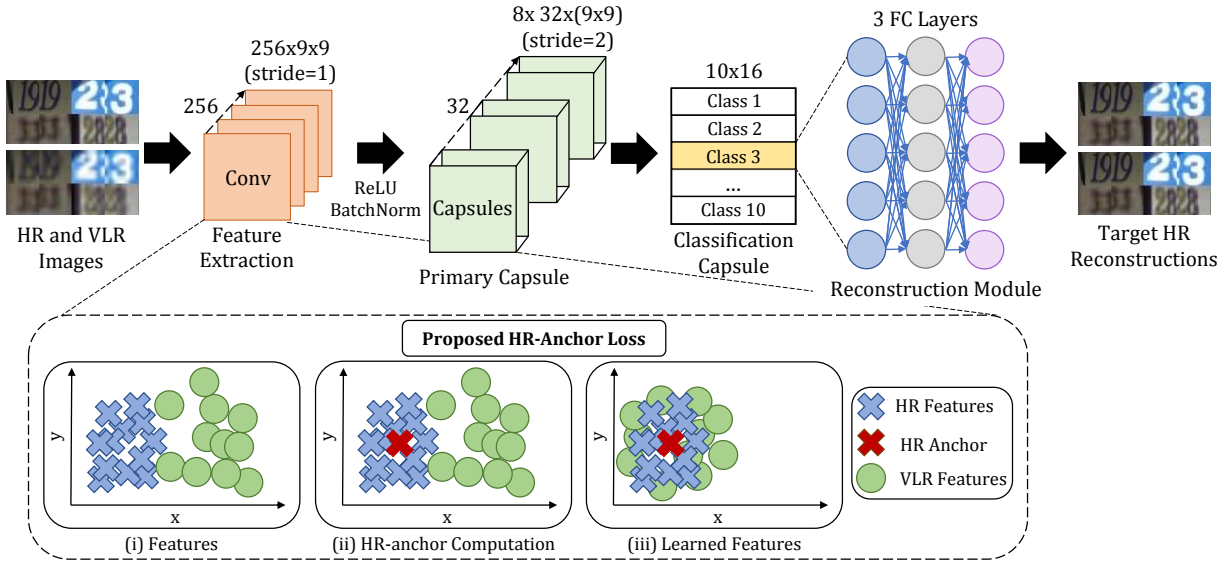



Figure 2-3: Architecture of the proposed Dual Directed Capsule Network (DirectCapsNet) for the SVHN dataset [104]. A diagrammatic representation of the HR-Anchor loss is presented for a given class. HR images are used to complement the features learned by the VLR recognition model by directing the model to learn discriminative and information rich features. The figure has been taken from the published manuscript [138].

ing synthetic data [60], and image classification [178]. Capsule networks encode the instantiation parameters for a given input, and thus present the potential of being the appropriate network for VLR image recognition.

### 2.3.2 Proposed DirectCapsNet

As shown in Figure 2-3, the proposed DirectCapsNet network can be broken down into three components: (i) input, (ii) feature extraction, and (iii) classification. At the time of training, the input consists of both HR and VLR samples. The feature extraction module consists of convolutional layers and the proposed HR-anchor loss, and the classification module consists of a capsule network coupled with the proposed targeted reconstruction loss. By enforcing dual direction via the proposed (i) HR-anchor loss and (ii) targeted reconstruction, the proposed DirectCapsNet focuses on learning meaningful feature-rich representations for VLR inputs, aided by the auxiliary HR samples. The loss function of the proposed DirectCapsNet is formulated as:

$$\mathcal{L}_{DirectCapsNet} = \mathcal{L}_{Margin} + \lambda_1 \mathcal{L}_{HR-anchor} + \lambda_2 \mathcal{L}_{T-Recon} \tag{2.2}$$

where, $\lambda_1$ and $\lambda_2$ are used to balance the weights of the HR-anchor and targeted reconstruction loss with respect to the margin loss. The margin loss introduces discriminability between classes, while the HR-anchor loss and targeted reconstruction loss enforce information-rich representations at the feature and classification level. At the time of testing, for a given VLR input, the class capsule with the highest length is chosen as the class of the given input. It is essential to note that, simulating the real world scenarios, DirectCapsNet utilizes the HR samples only at the time of training, and operates with a given VLR image during testing. As will be demonstrated in the remainder of this section, each component of the proposed model facilitates learning discriminative features for VLR recognition.

**Proposed HR-anchor Loss:** Input samples in Figure 2-3 are HR ($32 \times 32 \times 3$) and VLR ($8 \times 8 \times 3$ resolution upscaled to HR resolution) images from the SVHN dataset [104]. The limited information content in VLR images makes it difficult to extract discriminative information, often resulting in ineffective recognition, a phenomenon observed in humans as well [146]. The proposed HR-anchor loss addresses this challenge by pushing VLR features closer to their HR counter parts. This ensures learning of a discriminative space for VLR recognition, even with limited information. For an input $x^c$ belonging to class $c$, with features $f^{x^c}$ learned from the convolutional layers, the HR-anchor loss is formulated as:

$$\mathcal{L}_{HR-anchor} = \frac{1}{2}\big((1 - r^{x^c})\|f^{x^c} - \mathbb{A}^c\|_2^2 \; + \; r^{x^c} \; \|f^{x^c} - A^c\|_2^2\big) \tag{2.3}$$

where, $r^{x^c}$ is a binary variable to denote the resolution of the sample i.e., $r^{x^c} = 1$ for a HR sample, and $r^{x^c} = 0$ for a VLR sample. Since HR samples are only used during training, this information is readily available. $f^{x^c}$ refers to the features extracted from the convolutional layers in the feature module, $A^c$ and $\mathbb{A}^c$ both refer to the HR-anchor of class $c$, which is used to enhance the VLR representations. Specifically, $\mathbb{A}^c$ refers to the HR-anchor in a constant state, whereas $A^c$ represents the HR-anchor in a parameter form, which needs to be optimized. The HR-anchor of a particular class corresponds to the average feature vector of all HR samples belonging to that class. Given a VLR sample ($r^{x^c} = 0$), the first part of Equation 2.3 ($\|f^{x^c} - \mathbb{A}^c\|_2^2$) is active, where the HR-anchor of class $c$ assists the VLR feature $f^{x^c}$ to be closer to the anchor, thereby facilitating learning of discriminative features useful for classification. For a HR sample ($r^{x^c} = 1$), the second half of

Equation 2.3 ($\|f^{x^c} - A^c\|_2^2$) becomes active, where both the HR-anchor and features are updated.

The proposed HR-anchor loss is a combination of learning the HR-anchors and learning the VLR features closer to the HR feature space, in order to learn discriminative VLR features. The first term attempts to direct the VLR features towards the HR anchors, and the second term learns representative HR anchors from the HR features. It is important to note that there is no contribution of the VLR features in the anchor generation, since the HR anchors are constant in the first term. This ensures that the VLR features are directed towards the higher quality HR features, and not the other way round. Therefore, Equation 2.3 promotes the learning of informative VLR features with assistance from the HR samples.

**Proposed Targeted Reconstruction Loss:** The second form of direction is imposed via the targeted reconstruction loss (Figure 2-3) at the classification module (capsule network). The targeted reconstruction loss promotes learning similar classification capsules for HR and VLR samples. As explained previously, a capsule is a vector which encodes the instantiation parameters of the input sample [131]. For a given input, the activations of a capsule are termed as the activity vector. For reconstruction, only the activity vector of the target class is selected and used to reconstruct the input sample. For an input image $x^c$ belonging to class $c$, the reconstruction loss is mathematically formulated as:

$$\mathcal{L}_{Recon} = \frac{1}{2}\|x^c - g(v_c^{x^c})\|_2^2 \tag{2.4}$$

where, $v_c^{x^c}$ is the activity vector of the classification capsule of the $c^{th}$ class for the input $x^c$, and $g(.)$ refers to the reconstruction network. The reconstruction loss attempts to encode instantiation parameters that are able to explain the input image, and thus are able to reconstruct the input. Intuitively, we believe that the instantiation parameters of a HR sample and its corresponding VLR sample should be similar. Therefore, in order to incorporate a second level of direction, the targeted reconstruction loss is introduced in the proposed DirectCapsNet.

The targeted reconstruction loss enforces the HR counter-part of a VLR image at the output of the reconstruction network. Regardless of a HR or a VLR input, the reconstructed sample is forced as a HR image. For an input $x^c$, the targeted reconstruction loss can be written as:

$$\mathcal{L}_{T-Recon} = \|hr^{x^c} - g(v_c^{x^c})\|_2^2 \tag{2.5}$$

where, $hr^{x^c}$ is the HR image corresponding to the input HR/VLR sample and $v_c^{x^c}$ is the activity vector of the $c^{th}$ class. In case of a HR input image, Equation 2.5 ensures that the HR input is reconstructed at the output of the reconstruction network. For a VLR image, its HR counter-part is provided as the target of the reconstruction network. Since the reconstruction network operates on the final classification capsule, the targeted reconstruction loss pushes the HR and VLR samples to have a similar capsule activity vector, driven by the HR samples. Therefore, the reconstruction loss promotes learning similar capsule features for HR and VLR samples directly at the classification stage, by directing the model to reconstruct a HR sample from an extracted VLR feature.

Equations 2.3 and 2.5 are combined to update Equation 2.1 and the loss function of the proposed DirectCapsNet for an input $x^c$ (belonging to class $c$) is written as:

$$\begin{aligned}\mathcal{L}_{DirectCapsNet} = \sum_{k=1}^{K} \Big( T_k \max(0, m^+ - \|v_k^{x^c}\|)^2 + \lambda\,(1 - T_k)\ \max(0,\ \|v_k^{x^c}\| - m^-)^2 \Big) \\ + \frac{1}{2}\Big( \lambda_1 (1 - r^{x^c})\ \|f^{x^c} - \mathbb{A}^c\|_2^2 \ + \lambda_1 r^{x^c}\ \|f^{x^c} - A^c\|_2^2 \ + \lambda_2\ \|hr^{x^c} - g(v_c^{x^c})\|_2^2 \Big)\end{aligned} \tag{2.6}$$

### 2.3.3 Implementation Details

DirectCapsNet has been implemented in Python, using the PyTorch framework on the NVIDIA Tesla P-100 GPU. Adam optimizer [69] has been used for learning the model. The weight of the HR-anchor loss ($\lambda_1$ of Equation 2.6) is set to $1e-3$, and the weight of the targeted reconstruction loss ($\lambda_2$ of Equation 2.6) is set to $1e-5$. The positive and negative margins for the margin loss ($m^+$ and $m^-$ of Equation 2.1) are set to 0.9 and 0.1, respectively. As shown in Figure 2-3, for all the experiments, the DirectCapsNet model contains $n$ convolution layers, followed by two capsule layers. The HR-anchor loss is applied on the final convolution layer of the DirectCapsNet. The final capsule layer is connected to a reconstruction network of three fully connected layers. For cases where the HR samples are larger than $96 \times 96$, three convolutional layers with $[16, 32, 128]$ filters are used with a batch size of 32 samples. In cases where the HR samples are smaller, a convolutional layer with 128 filters is used with a batch size of 100 samples. $ReLU$ activation function is used between the convolutional layers along with batch normalization [58]. All models have been trained from scratch and no pre-trained networks have been used.

## 2.4 Experiments and Protocols

The proposed DirectCapsNet has been evaluated for three very low resolution (VLR) recognition problems: (i) VLR digit recognition, (ii) VLR face recognition, and (iii) unconstrained VLR face recognition. Details regarding the dataset and protocols for each case study are as follows:

**Case study 1 - VLR Digit Recognition:** The Street View House Numbers (SVHN) dataset [104] has been used for VLR digit recognition. The dataset contains real-world images of digits in the range $[0 - 9]$. Pre-defined benchmark protocol has been used for the given 10-class problem, wherein 73,257 digits are used for training and 26,032 digits are used for testing. For VLR recognition, consistent with the existing protocol [170], $32 \times 32$ HR images are used, and $8 \times 8$ VLR images are used. Results are reported in terms of the top-1 and top-5 accuracies.

**Case study 2 - VLR Face Recognition:** VLR face recognition has direct applicability in scenarios of image tagging or situations where multiple people are captured in a single image. For this particular case-study, experiments have been performed on the CMU Multi-PIE dataset [44] which simulates a constrained setting. Consistent with the existing protocol [142], 237 subjects are used. One image per subject is added to the training set/gallery which consists of the HR images, and one image per subject is added to the testing set/probe (VLR). The HR images are of $96 \times 96$ resolution and the VLR images are of $8 \times 8$ and $16 \times 16$, respectively. Results are reported using the rank-1 identification accuracy.

**Case study 3 - Unconstrained VLR Face Recognition:** Unconstrained VLR face recognition has wide applicability in surveillance scenarios, where the VLR face image often contains other variations such as pose, illumination, and occlusion. Experiments have been performed on two datasets: (a) UnConstrained College Students (UCCS) dataset [133] for an unconstrained surveillance setting and (b) CMU Multi-PIE dataset [44] with pose and illumination variations for a semi-constrained setting.

The UCCS dataset contains images of college students, captured using a long-range high-resolution surveillance camera kept at a standoff distance of 100 to 150 meters. The images show students walking around the campus, between classes. The large standoff distance and unconstrained nature of the data simulates real world surveillance settings. The dataset contains a labeled subset of 1732 identities. Consistent with the existing protocol [37, 170], a subset containing the top 180

identities (in terms of the number of images) is used for evaluation. As per the protocol, each subject's images are divided into a $4:1$ ratio corresponding to training:testing. The VLR images are of $16 \times 16$ resolution, whereas the HR images are of $80 \times 80$ pixels.

As described above, CMU Multi-PIE dataset [44] contains images with pose, expression, and illumination variations. As per the existing protocol [102], in this case-study face recognition is performed across pose and illumination variations for VLR images. Images pertaining to 50 subjects are used for training and images of the remaining subjects form the test set. In our experiments, we do not utilize the training set and only use the gallery images of the test set in order to train the proposed DirectCapsNet model. The gallery comprises of the frontal images (used for training the proposed model), and the probe (test set) are images having a different pose ('05_0' of the dataset) and illumination. Experiments are performed across five different pairs of illumination conditions and average rank-1 identification accuracy has been reported. Consistent with [102], the HR images are of $36 \times 30$ resolution, while VLR images have resolution of $18 \times 15$, $15 \times 12$, $12 \times 10$, and $10 \times 9$.

Figure 2-2 presents some HR and VLR images from the datasets used in the three case-studies. Bicubic interpolation is used for conversion from HR to VLR and vice-versa. At the time of training, the HR and VLR pairs are used for the targeted reconstruction loss. Data augmentation is applied by introducing brightness variations, flipping along the y-axis, and random crops. At the time of testing, only the VLR image is provided for classification.

## 2.5 Results and Analysis

Tables 2.1 - 2.3 and Figures 2-4 - 2-6 present the results for the three case-studies: (i) VLR digit recognition, (ii) VLR face recognition, and (iii) unconstrained VLR face recognition. Analysis of the proposed DirectCapsNet has also been performed in order to demonstrate the effectiveness of each component. Since existing protocols have been used for analysis, results have directly been reported from the respective publications.

**Ablation Study and Analysis of DirectCapsNet:** Experiments have been performed on the SVHN dataset to analyze each component of the proposed DirectCapsNet, and motivate their inclusion in the final model. As observed from Table 2.1, the native CapsNet model (having the margin

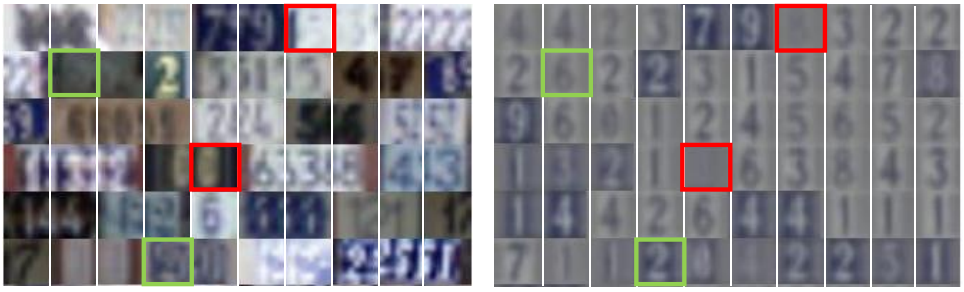
Table 2.1: Top-1 and top-5 accuracy (%) on the SVHN dataset [104] for VLR digit recognition ($8 \times 8$).

| | **Algorithm** | **Accuracy (%)** | |
|---|---|---|---|
| | | Top-1 | Top-5 |
| | CNN (VLR) (2016) [170] | 45.29 | 66.78 |
| | RPC Nets (2016) [170] | 56.98 | 70.82 |
| Proposed | CapsNet (HR) | 77.82 | 87.86 |
| | CapsNet (VLR) | 79.19 | 88.89 |
| | DirectCapsNet - (HR-anchor Loss) | 82.42 | 90.15 |
| | DirectCapsNet - (Targeted Recon.) | 81.95 | 90.35 |
| | **Proposed DirectCapsNet** | **84.51** | **91.20** |

loss) when trained on VLR images (CapsNet (VLR)) attains the top-1 classification accuracy of 79.19%, which demonstrates large improvement over the native CNN architecture (45.29%) [170]. The improved performance promotes the usage of capsule networks for the task of VLR recognition. Consistent with literature [131], it is our belief that since CapsNet attempts to encode the instantiation parameters of the data, it results in learning features invariant to minor variations, a desirable property of a robust VLR recognition module.

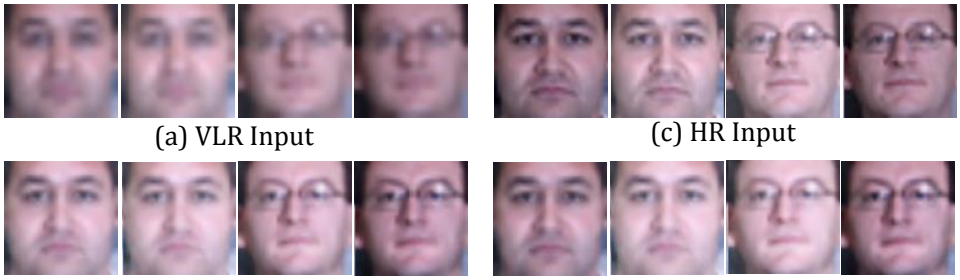
Further, in order to reaffirm the necessity of a VLR recognition model, a CapsNet with the same architecture is trained on HR images only. In this case, the model does not see any VLR images at the time of training and is evaluated on VLR test images. As can be observed, the CapsNet (HR) achieves a classification accuracy of 77.82%, thus reaffirming the need to develop dedicated VLR recognition networks or utilize task-specific information while training. We also performed the McNemar test [96] and achieved statistical difference at a confidence interval (C.I.) of 99% ($p$-value$<$0.01) between the proposed DirectCapsNet and CapsNet. Table 2.1 can also be analyzed to understand the effect of each component of the proposed DirectCapsNet model. Upon removing the HR-anchor loss from the DirectCapsNet model, top-1 accuracy of 82.42% is achieved, whereas, removal of the targeted reconstruction loss results in a top-1 accuracy of 81.95%. Both these models demonstrate poor performance as compared to the proposed DirectCapsNet model, thus supporting the inclusion of the HR-anchor loss, targeted reconstruction loss, and capsules in the DirectCapsNet model.

**Case study 1 - VLR Digit Classification:** Table 2.1 presents the top-1 and top-5 classification



(a) VLR Input (b) Reconstructions

Figure 2-4: Sample reconstructions obtained on the SVHN dataset from VLR input. DirectCapsNet is able to reconstruct digits where limited information content is available (e.g., green boxes), however it also fails to correctly reconstruct some challenging cases (e.g., red boxes).



(a) VLR Input (c) HR Input

(b) Reconstruction for VLR Input (d) Reconstruction for HR Input

Figure 2-5: Sample reconstructions obtained from the proposed DirectCapsNet model on the CMU Multi-PIE dataset. For the same class, DirectCapsNet is able to project VLR and HR samples onto a similar target, suggesting robust resolution-invariant feature representations.

accuracy for the SVHN dataset of the proposed DirectCapsNet and comparison with other techniques. The proposed DirectCapsNet model achieves top-1 accuracy of 84.51% and top-5 accuracy of 91.20%. DirectCapsNet demonstrates an improvement of over 27% at top-1 with respect to the state-of-the-art results of Robust Partially Coupled Networks (RPC Nets) [170], which is a CNN based framework to learn partial shared weights for VLR and HR samples, and partial independent weights for the two. The superior performance of the proposed DirectCapsNet model motivates its usage for VLR recognition. Figure 2-4 presents sample reconstructions obtained from the DirectCapsNet for $8 \times 8$ VLR samples. It is motivating to note that the DirectCapsNet model is able to reconstruct the digits for the input samples, which motivates the inclusion of the targeted reconstruction loss. Similar reconstructions are obtained for samples of the same class, which demonstrate the effectiveness of the HR-anchor loss for increasing the intra-class similarity between features.

Table 2.2: Rank-1 accuracy (%) for VLR recognition on the CMU Multi-PIE dataset [44]. The HR images are of $96 \times 96$ resolution.

| **Algorithm** | **Accuracy (%)** | |
|---|---|---|
| | $8 \times 8$ | $16 \times 16$ |
| Original + COTS (2018) [142] | 0.0 | 0.0 |
| Bicubic Interp. + COTS (2018) [142] | 0.1 | 1.1 |
| SHSR (Synthesis + COTS) (2018) [142] | 82.6 | 91.1 |
| **Proposed DirectCapsNet** | **94.5** | **97.4** |

Table 2.3: Rank-1 accuracy (%) on the UCCS dataset [133] for VLR face recognition ($16 \times 16$). The HR images are of $80 \times 80$ resolution.

| **Algorithm** | **Acc. (%)** |
|---|---|
| Robust Partially Coupled Nets (2016) [170] | 59.03 |
| Selective Knowledge Distillation (2019) [37] | 67.25 |
| LMSoftmax for VLR (2019) [81] | 64.90 |
| L2Softmax for VLR (2019) [81] | 85.00 |
| Centerloss for VLR (2019) [81] | 93.40 |
| **Proposed DirectCapsNet** | **95.81** |

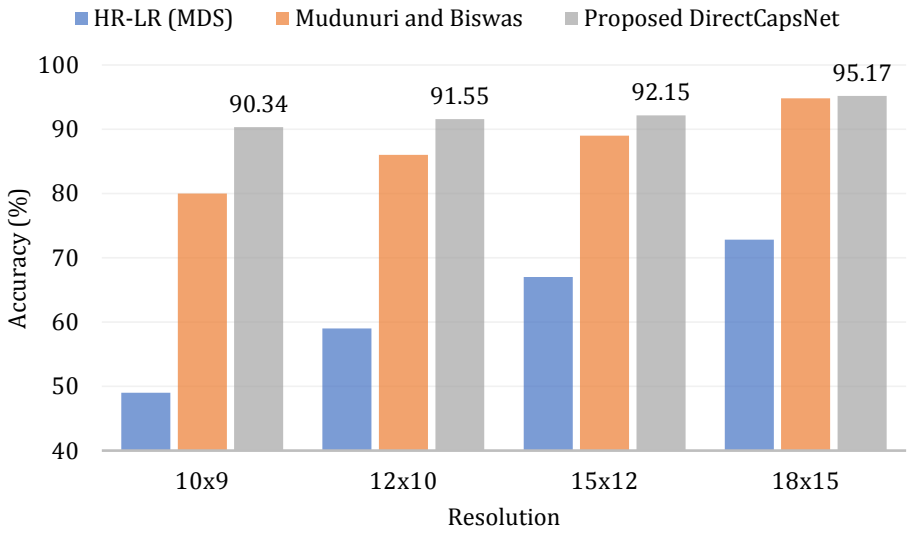



Figure 2-6: Performance of the proposed DirectCapsNet for varying resolutions of VLR face recognition with pose and illumination variations. The HR resolution was fixed to $36 \times 30$ pixels. Comparison has been shown with HR-LR (MDS) [13] and Mudunuri and Biswas [102].

**Case study 2 - VLR Face Recognition:** Table 2.2 presents the rank-1 identification (or top-1 recognition) accuracy for two protocols of VLR face recognition. The proposed DirectCapsNet model achieves an accuracy of 94.5% and 97.4% for $8 \times 8$ and $16 \times 16$ VLR images, while

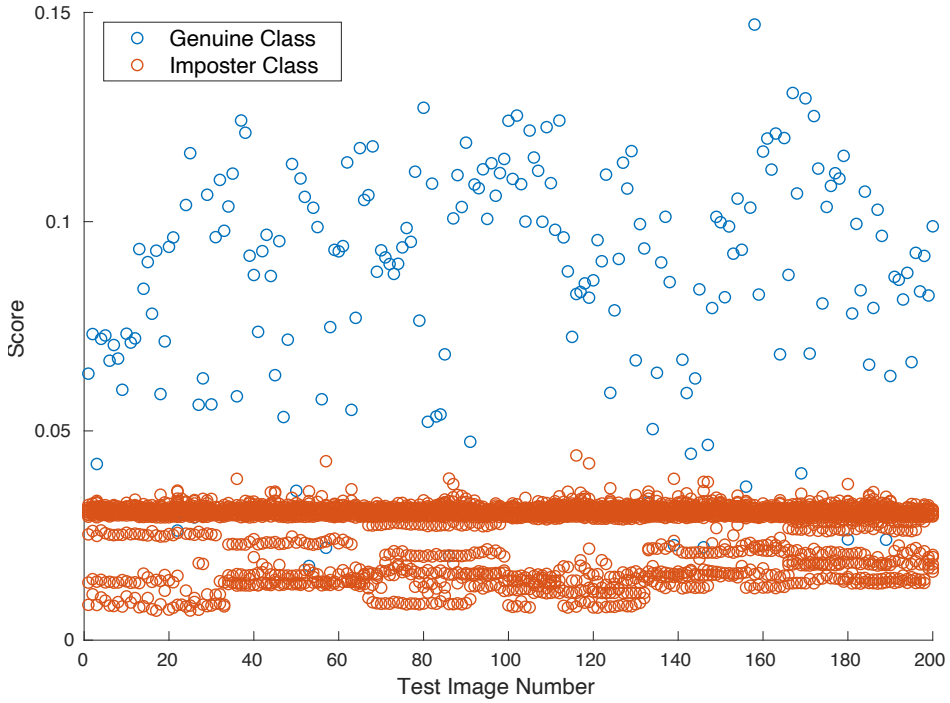



Figure 2-7: Scores obtained by the proposed DirectCapsNet for VLR recognition on some samples of the UCCS dataset. Each test image has one genuine score (correct class) and 179 imposter scores (incorrect class).

having the HR auxiliary images as $96 \times 96$ (Table 2.2) on the constrained CMU Multi-PIE dataset. DirectCapsNet demonstrates an improvement of almost 12% as compared to the state-of-the-art (Synthesis via Hierarchical Sparse Representations (SHSR)) [142] for $8 \times 8$ resolution images. Figure 2-5 presents sample VLR and HR face images, along with the reconstructions obtained from the DirectCapsNet. The proposed model is able to reconstruct faces belonging to the same subject onto a similar target, suggesting high within-class similarity. Both VLR and HR samples are reconstructed as similar images, which reinstates the benefit of the targeted reconstruction and HR-anchor loss.

**Case study 3 - Unconstrained VLR Face Recognition:** Table 2.3 and Figure 2-6 present the rank-1 identification (or top-1 recognition) accuracy for unconstrained VLR face recognition. As shown in Table 2.3, on the UCCS dataset, DirectCapsNet model achieves a rank-1 accuracy of 95.81% demonstrating an improvement of almost 2.5% over the state-of-the-art network and almost 10% from the current second best [81]. Comparison has also been performed with the recently proposed

large-margin softmax (LMSoftmax), $l_2$-constrained softmax (L2Softmax), and center-loss based VLR recognition systems [81]. The improved performance of the proposed DirectCapsNet over metric learning techniques demonstrates the benefit of incorporating auxiliary HR information to provide direction while training with the proposed dual directed loss functions. Figure 2-7 presents the scores obtained on samples of the UCCS dataset by the DirectCaspNet model. The scores correspond to the length of the activity vectors of the capsules used for classification. Figure 2-7 suggests that the model generates a high score for the correct class and a small score for the other classes, which promotes separability, resulting in high recognition performance.

Similar performance is obtained on the CMU Multi-PIE dataset (Figure 2-6) with pose and illumination variations, where the proposed DirectCapsNet achieves an average recognition performance of 95.17%, demonstrating an improvement of around 1.64% from the current state-of-the-art algorithm [102]. Figure 2-6 demonstrates that the proposed DirectCapsNet does not suffer a major decrease in accuracy as other techniques with reducing the resolution. The model achieves the recognition accuracy of 92.15% and 90.34% for $15 \times 12$ and $10 \times 9$, respectively, whereas, the second best performing model [102] shows a drop of almost 9% between the two resolutions. Improved recognition performance across multiple very low resolutions motivates the applicability of the proposed DirectCapsNet model for real world scenarios.

## 2.6 Summary

Existing research has primarily focused on high resolution and low resolution image recognition; however, the problem of VLR recognition has received limited attention. VLR recognition, an arduous problem with wide applicability in real world scenarios, suffers from the primary challenge of low information content. This research presents a novel Dual Directed Capsule Network (DirectCapsNet) for VLR recognition. The DirectCapsNet combines the margin loss for classification with the proposed HR-anchor loss and the targeted reconstruction loss for enhancing the VLR features. HR images are used during training as 'auxiliary' data to complement the VLR feature learning. Experimental results on VLR digit recognition (SVHN database) and constrained/unconstrained VLR face recognition (CMU Multi-PIE and UCCS databases) demonstrate the efficacy of the proposed model, and promote its usability for different VLR tasks. This research

has been published in the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, and all the figures have been taken from the published manuscript [138].

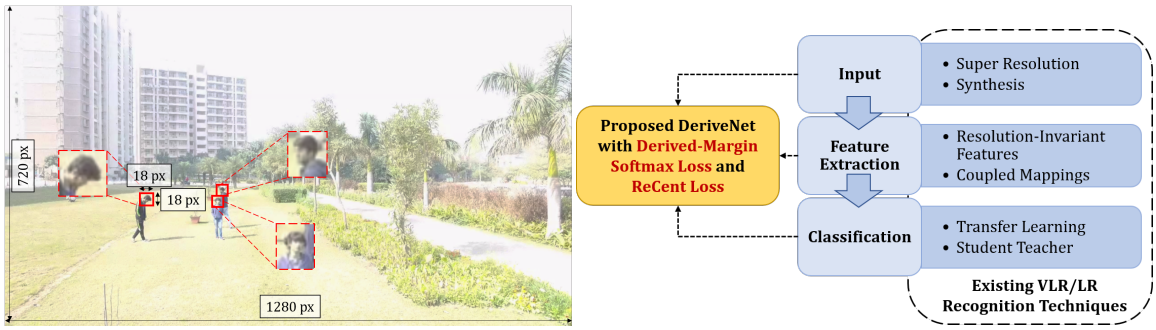
# Chapter 3

# DeriveNet for (Very) Low Resolution Image Classification

## 3.1 Introduction

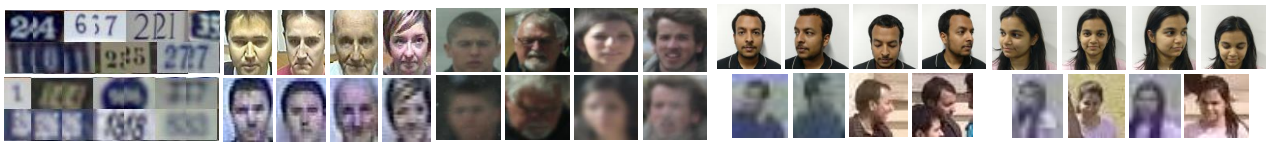
In real world scenarios, sometimes even in high resolution (HR) images, the region of interest (ROI) is small due to the large stand-off distance or large field of view of the acquisition device. As shown in Figure 3-1(a), the resultant ROI captured is of low resolution (LR) or very low resolution (VLR), even though the size of the complete image is large. Generally, in the literature, LR classification refers to identifying image regions with a resolution of $32 \times 32$ to $16 \times 16$ pixels. When the resolution of the region of interest is even lower than $16 \times 16$ pixels, the task is referred to as VLR recognition/classification [201].

In the literature, majority of the research has focused on high resolution image classification tasks for various applications, with limited focus on VLR/LR recognition. As shown in Figure 3-2, VLR/LR images contain limited information content as compared to the HR images. The lower resolution of images and limited interpretive information makes it difficult to perform automated classification/recognition. VLR classification often also faces the challenge of cross-resolution matching, where the query/probe is of a (very) LR, while the gallery is often of high resolution. Further, the variations observed across the HR and VLR/LR images make it challenging to directly utilize the models trained on good quality HR images for VLR/LR images. The arduous nature of the problem and the ubiquitous applicability of the task demands dedicated research attention.




(a) Sample frame from the DroneSURF dataset [64]. (b) Existing techniques and proposed DeriveNet model.

Figure 3-1: (a) VLR/LR face images captured in an unconstrained environment often contain limited information content, rendering face recognition challenging. (b) Existing techniques for VLR/LR image classification often focus only on the image, feature, or classification space. The proposed DeriveNet models both the image and classification space for VLR/LR recognition, while implicitly learning effective features.



(a) SVHN Dataset (b) Scface Dataset (c) UCCS Dataset (d) DroneSURF Dataset

Figure 3-2: Sample HR and VLR/LR images from different datasets: (a)/(b-c) VLR digit/face recognition and (d) VLR/LR face recognition in drone-shot videos. The top and the bottom row contains the HR and VLR/LR images, respectively.

Learning a (very) low resolution classification model suffers from the challenge of limited information content in the input images, thus making it difficult to extract useful features for classification. The problem is further amplified due to the low inter-class variations observed from the samples. Therefore, most of the existing algorithms for VLR/LR classification use some auxiliary HR data along with the LR images for learning an effective model (also used by the proposed model), while focusing either on the image space (super-resolution, synthesis or down-sampling), feature space (feature transformations), or the classifier space (Figure 3-1(b)). As discussed in Section 3.2, most of the algorithms also pose the requirement of mated (corresponding) HR-LR pairs, which is often not available in real world scenarios. Further, despite the recent advances for VLR/LR recognition tasks, there is still a wide gap between the state-of-the-art performance on VLR/LR images and HR images. For example, very high accuracy has been obtained on the challenging high resolution Labeled Faces in the Wild dataset [57], whereas the performance on

VLR/LR face images from the DroneSURF dataset [64] is less than 20% [5]. Thus, in order to address the above limitations, this research proposes a novel DeriveNet model for VLR/LR images, which focuses on all three aspects of the classification pipeline. DeriveNet operates at the intersection of the image space and the classification space, while implicitly learning efficient features, using LR images and auxiliary HR images during training. The contributions of this research are:

- We propose a novel *DeriveNet* model for VLR/LR image classification. The proposed model utilizes two novel loss functions: (i) Derived-Margin softmax loss and (ii) Reconstruction-Center (ReCent) loss. Both loss functions facilitate the DeriveNet model to address the challenge of poor feature extraction from VLR samples and focuses on learning more efficient classification boundaries by utilizing class-specific domain knowledge.

- A novel Derived-Margin softmax loss has been proposed which utilizes a *derived margin* to model the class variations by promoting larger distance between classes having higher similarity. Instead of an arbitrary margin value, the derived margin is obtained via the proposed ReCent loss which uses auxiliary HR samples (at least one sample per class) during training to approximate the class variations of VLR/LR samples while learning similar features for VLR/LR/HR images.

- The DeriveNet model has been trained using a novel *Multi-resolution Pyramid based data augmentation* technique which enables it to learn from different resolutions during training.

- The efficacy of the proposed model is demonstrated via experiments on five datasets for different applications of (i) VLR/LR face recognition, (ii) VLR digit classification, and (iii) VLR/LR face recognition in drone-shot videos. Training has been performed using a novel Multi-resolution Pyramid based data augmentation technique which demonstrates improved performance. For example, the proposed model achieves over 97% rank-1 face recognition performance on the UCCS dataset [133] and 84% on the SCface database [42].

## 3.2 Related Work

In the literature, existing techniques for VLR/LR recognition often utilize HR information for learning improved classification models, and can be categorized into one of the following cat-

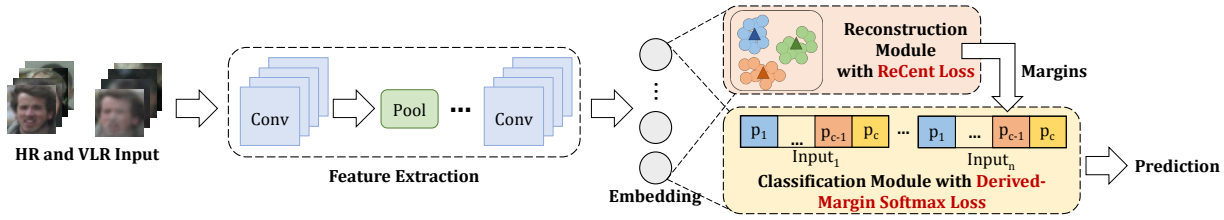



Figure 3-3: Illustration of the training phase of the proposed DeriveNet model for VLR face recognition. HR and VLR images are used as input, followed by a convolutional neural network based feature extractor. The extracted embedding are provided as input to two modules: (i) the reconstruction module and the (ii) classification module. During testing, a given VLR input image is passed through the feature extractor, followed by the classification module for obtaining the predicted class. The figure has been taken from the published manuscript [139].

egories: (i) image-level, (ii) feature-level, or (iii) classifier-level. At the image-level, super-resolution or synthesis [142, 185] has been explored, with the initial research focusing on enhancing the visual quality only, followed by techniques which optimize both the recognition performance and the visual quality [54, 184, 188]. For example, recently, Grm *et al.* [43] proposed a cascaded super-resolution network, along with an ensemble of face recognition models as identity priors, while Kazemi *et al.* [66] proposed utilizing a multi-scale generative adversarial network for the same task. Generally, most of these algorithms often require corresponding HR-LR pairs for training, and present the additional burden of enhancing the visual quality, which is often not a required output for a classification system.

Researchers have also proposed VLR/LR algorithms which incorporate the HR information at the feature level or at the classifier level. Ouyang *et al.* [110] presented a comprehensive survey of techniques proposed specifically for low resolution face recognition, and Aghdam *et al.* [3] analyzed the effect of different factors on low resolution face recognition. Wang *et al.* [170] proposed Robust Partially Coupled Networks for learning a classification model for VLR images. Shared and individual weights were learned from the HR and VLR images for learning an effective classifier. Lu *et al.* [88] proposed Coupled Residual Networks for learning a LR face recognition model, where feature-level differences were minimized between the HR and LR images. Recently, Li *et al.* [81] presented a comprehensive analysis of recent techniques on VLR/LR face recognition tasks. Ge *et al.* [37] proposed learning a student-teacher network for utilizing a HR-trained model for VLR/LR face recognition. Singh *et al.* [138] proposed Dual Directed Capsule Network for VLR classification, where information from the HR images are used to direct

a VLR recognition capsule network. Algorithms have also been proposed which focus on down-sampling the HR images to match the resolution of the LR/VLR sample, followed by matching at the lower resolution [95, 193]. Further, recently several drone-based datasets have also been released [27, 64, 72, 198] containing videos/frames captured from the drone. Most of such datasets focus on pedestrian/vehicle detection or tracking, with limited focus on face/object recognition. Other than person/object classification, researchers have also focused on low resolution activity recognition [129, 130].

## 3.3 DeriveNet for (V)LR Recognition

As observed from Figure 3-2, VLR/LR samples often contain limited interpretive information. This often leads to high inter-class similarity, challenging discriminative feature extraction, and poor classification performance. In order to address the above limitations, this research proposes a novel DeriveNet model for VLR/LR classification. The DeriveNet model learns effective class boundaries based on a *derived margin* which enforces larger distance between similar classes. As shown in Figure 3-3, the DeriveNet model contains (i) a feature extraction module, (ii) a reconstruction module, and (iii) a classification module.

- The feature extraction module is used for extracting feature embedding from the given input image. Training is performed using the VLR/LR images and some auxiliary HR samples (at least one per class). During training, the embedding is provided to the (a) reconstruction module for constructing a corresponding HR image, and (b) the classification module for learning an effective classifier (Figure 3-4).

- The reconstruction module constructs a HR image for the given VLR/LR/HR input (using the same network), and is optimized via the *Reconstruction-Center loss (ReCent loss)*. The ReCent loss serves the dual purpose of constructing HR images and learning HR centers for each class which are used for deriving the inter-class distance or margin for the classification module.

- The classification module utilizes the feature embedding and the margin values for learning effective class boundaries. It is optimized via the proposed *Derived-Margin Softmax loss*

which promotes larger distance between similar classes.

For an input feature $x_i$ of class $y_i$, and a corresponding HR image ($hr_i$), the ReCent loss and the Derived-Margin softmax loss are jointly used for the DeriveNet model (for a $C$ class problem containing $N$ training samples) as follows:

$$\mathcal{L}_{DeriveNet} = \frac{1}{N}\sum_{i=1}^{N}\Bigg(\underbrace{-\log\frac{\exp^{\|W_{y_i}\|\|x_i\|\cos\theta_{y_i}}}{\sum_{j=1}^{C}\exp^{\|W_j\|\|x_i\|\cos\theta_j+\mathcal{D}(C_{y_i},C_j)}}}_{\mathcal{L}_{D-Margin}}\underbrace{+\lambda(\|hr_i-g(x_i)\|_2^2+\|g(x_i)-C_{y_i}\|_2^2)}_{\mathcal{L}_{ReCent}}\Bigg) \tag{3.1}$$

where, $W_j$ and $C_j$ refer to the classification weight vector and HR class center for the $j^{th}$ class, respectively, $g(.)$ refers to the reconstruction module, $\mathcal{D}(.)$ refers to a similarity function, and $\lambda$ refers to the weight of the ReCent loss. The first term (Derived-Margin softmax loss) learns an efficient VLR/LR classifier with explicit focus on the inter-class variations by utilizing a margin penalty ($\mathcal{D}(C_{y_i}, C_j)$: inter-class similarity between class $y_i$ and $j$). Here, $\mathcal{D}(.)$ is modeled as a similarity function, such that a larger penalty is applied to more similar classes. The margin parameter is *derived* via the ReCent loss: the second term is used for projecting the given VLR/LR/HR embedding onto the reconstruction space, and the final term promotes learning compact HR constructions for each class, along the class centers. The distance between the learned HR class centers are used for obtaining the margin parameter between the two classes ($\mathcal{D}(C_{y_i}, C_j)$). Details of each component are as follows.

### 3.3.1 Derived-Margin Softmax Loss ($\mathcal{L}_{D-Margin}$)

The softmax function has widely been used in conjunction with the cross-entropy loss for classification tasks, in various deep learning models. Similar to recent literature [83], in this research, *the softmax loss has been defined as a combination of the softmax function and the cross-entropy loss on the final fully-connected layer*. The softmax loss attempts to learn a discriminative classifier which provides a larger score for the correct class of the input feature. The class scores are often the activations of the final fully-connected layer ($f_j = W_j^T x_i$; score of input ($x_i$) for class $j$). Liu *et al.* [83] formulated the softmax loss as a function of the angle between $x_i$ (input feature) and $W_j$ (weight matrix of the $j^{th}$ class) by reformulating the class score as a Cosine function. The updated

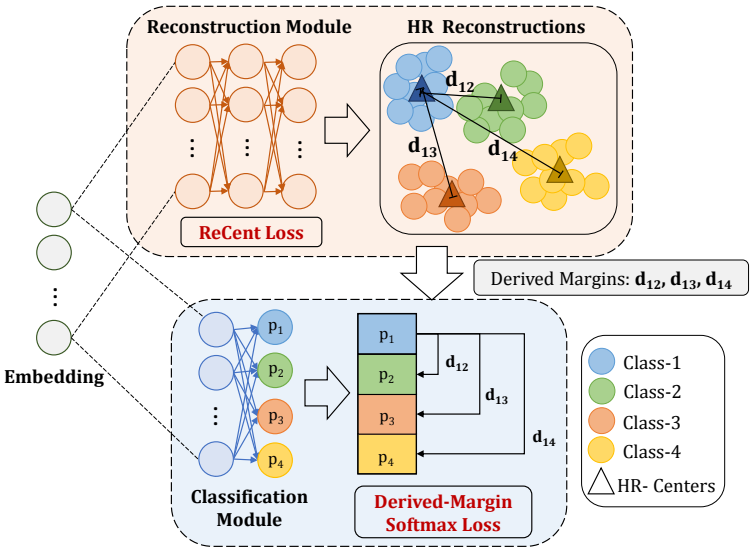



Figure 3-4: Diagrammatic representation of the DeriveNet model for a four class problem, with an input of class-1. The extracted embeddings are provided as input to the reconstruction and classification modules. The distance between the centers of the reconstructions are provided as margins to the proposed Derived-Margin softmax loss for learning a VLR/LR classifier.

Softmax loss is as follows:

$$\mathcal{L}_{Softmax} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp^{\|W_{y_i}\|\|x_i\|\cos\theta_{y_i}}}{\sum_{j=1}^{C}\exp^{\|W_j\|\|x_i\|\cos\theta_j}} \tag{3.2}$$

The above interpretation has been used for proposing different loss functions [20, 164], specifically for the task of high resolution face recognition. Almost all the techniques focus on incorporating an arbitrary margin parameter in the numerator (correct class component) of the softmax loss (Equation (3.2)), with no focus on the other-class components (denominator). As elaborated in the remainder of this Section, this interpretation of the softmax loss has been used for proposing the Derived-Margin softmax loss.

While the softmax loss (Equation (3.2)) can directly be applied for the task of VLR/LR classification tasks, it does not explicitly model the challenge of limited information content in VLR/LR images, often resulting in low inter-class similarity. To this effect, this research proposes a novel Derived-Margin softmax loss which utilizes class-specific domain knowledge for *deriving* the margins. For an input feature $x_i$ of class $y_i$, mathematically, the proposed Derived-Margin softmax loss

is given as:

$$\mathcal{L}_{D-Margin} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp^{\|W_{y_i}\|\|x_i\|\cos\theta_{y_i}}}{\sum_{j=1}^{C}\exp^{\|W_j\|\|x_i\|\cos\theta_j + d_{y_i j}}} \tag{3.3}$$

where, $d_{y_i,j}$ is the margin or penalty parameter for the $j^{th}$ class from the $y_i^{th}$ class. Here, the class-specific margins attempt to encode the similarity between different classes, thus modeling the inter-class variations. For example, in a very simplistic scenario of digit classification, one would expect a higher margin or penalty for similar looking classes such as $1$ and $7$, or $5$ and $6$, on the other hand, a lower margin or penalty can be applied on classes that appear easily distinguishable, such as $1$ and $0$, or $6$ and $7$. The Derived-Margin softmax loss enforces such variable margins or penalties by incorporating the derived margins ($d_{y_i j}$), thereby promoting a wider difference between the scores of similar classes.

Geometrically, for a two class problem (1 or 2), softmax requires ($W_1^T x > W_2^T x$) for a sample $x$ of class-1. DeriveNet forces the model to have ($W_1^T x > W_2^T x + d_{12}$), i.e, it introduces a penalty or margin between the scores of the two classes. For a multi-class model, DeriveNet enforces a margin between the scores of the correct class and all other classes, thus learning an effective classifier. Since the margin parameter encodes the inter-class variations, it is zero for the same class i.e., $d_{y_i,j} = 0$, when $j = y_i$. The ReCent loss is used for modeling the class variations and computing the derived margins.

### 3.3.2 Reconstruction-Center Loss ($\mathcal{L}_{ReCent}$)

As shown in Figure 3-4, the embeddings extracted from the input data are provided as input to the reconstruction and classification modules. The reconstruction module is a small fully-connected network used to construct a high resolution (HR) image from a given embedding. The reconstruction space is further utilized to incorporate intra-class similarity and derive class-specific margins for the Derived-Margin softmax loss. In order to achieve this, the ReCent loss is applied on the reconstruction module, which focuses on reducing the variability in the class-specific reconstructions [173]. For a feature $x_i$ and its HR image ($hr_i$), the ReCent loss is formulated as:

$$\mathcal{L}_{ReCent} = \|hr_i - g(x_i)\|_2^2 + \|g(x_i) - C_{y_i}\|_2^2 \tag{3.4}$$

where, $g(.)$ refers to the reconstruction module used for constructing a HR image from the given feature embedding ($x_i$), and $C_{y_i}$ refers to the center of the constructed HR images of class $y_i$. For a given class with $n$ samples, the center can be viewed as the mean of the constructed HR images for that class. For each class, the center corresponds to a $k \times 1$ parameter, where $k$ is the dimension of the flattened HR image, and is updated in each mini-batch via gradient descent. If some classes are not present in a mini-batch, the corresponding centers are not updated. The ReCent loss extends the traditional Center Loss [173] by constructing a HR image from the penultimate layer of a given VLR/LR/HR input and learning class-specific centers in the constructed HR space. The learned centers are then used for calculating the derived margins which are used by the novel Derived-Margin softmax loss for learning an effective classifier. The reconstruction module promotes meaningful VLR/LR features capable of constructing a HR image, while learning class-wise compact HR constructions (closer to the respective class centers). Since the class center is learned during the model training (and is viewed as the mean of each classâĂŹs samples), the center is updated to reflect the variations seen in the reconstructed HR samples of a given class. Therefore, the two components jointly approximate the class variations observed in the HR samples. Further, since the original LR/VLR images do not exhibit the original inter-class structure observed in the HR samples, the learned class centers are used for deriving the margins. Since the effect of the ReCent loss is back-propagated to the features used for classification, the ReCent loss promotes learning features representative of both the VLR and HR images, while being effective for classification. Thus, the aim of the ReCent loss is to approximate the class variations for VLR/LR samples and not perform super-resolution or synthesis. It is important to note that the ReCent loss requires a small pool of HR images during training (at least one per class). Further, in the real world scenario of no mated or corresponding VLR/LR and HR samples, the ReCent loss can be optimized by combining a VLR sample with any HR sample of the same class.

The proposed DeriveNet model for VLR/LR classification incorporates the distance between the HR class centers of the reconstructed space (obtained via the ReCent loss) into the Derived-Margin softmax loss. The distance between the centers of two reconstructed classes provide an estimate of the similarity between their samples or the inter-class variations. Mathematically,

between centers $C_i$ and $C_j$, the *derived margin* for Equation (3.3) is given as:

$$d_{ij} = \mathcal{D}(C_i, C_j) \tag{3.5}$$

where, $\mathcal{D}(.)$ is a distance function applied on $C_i$ and $C_j$. A small distance value corresponds to high similarity, while a larger distance refers to low similarity between the two classes. As demonstrated in Figure 3-4 the Derived-Margin softmax loss and the ReCent loss are combined (Equation 3.1) for learning an efficient DeriveNet model.

### 3.3.3 Training DeriveNet with Multi-Resolution Pyramid based Data Augmentation

During the training of the DeriveNet model, we propose to use multi-resolution pyramid based data augmentation. In this approach, high resolution images are down-sampled to generate corresponding low resolution images from the following resolutions (in a pyramid manner): [$8 \times 8$, $16 \times 16$, $24 \times 24$, $32 \times 32$, $48 \times 48$, and $64 \times 64$]. During training, for each epoch, randomly three lower resolution versions of each image are selected and used as data augmentation. This enables the model to learn from different resolutions in a time-efficient manner, while preventing redundancy in the training samples. Applying multi-resolution pyramid based data augmentation during pre-training enables the model to process large number of low-resolution samples during training, thus resulting in better feature learning for the task of VLR/LR classification. In the literature, Massoli *et al.* [95] also used a large-scale varying resolution dataset during training, however, the algorithm generated LR images on randomly chosen resolutions (range of [8, 256] pixels), whereas the pyramid based augmentation generates samples from a fixed step of resolutions.

### 3.3.4 Implementation Details

As demonstrated in Figure 3-3, DeriveNet contains a feature extraction module, a reconstruction module, and a classification module. If the input image is of size greater than $96 \times 96$, the pre-trained deep CNN architecture presented by Wu *et al.* [177] containing 17 convolutional layers and 10 Max-Feature-Map layers, is used as the feature extractor, which is fine-tuned with the training

data (details provided in Section 3.4). For other cases, the feature extraction module consists of a two convolutional layers with $[512, 512]$ filters, along with ReLU activation and adaptive pooling. For all the experiments, the reconstruction module consists of three fully connected layers ($[1024, 2048, k]$, where $k$ is the input dimension), and the classification module consists of two fully connected layers ($[128, n]$, where $n$ is the number of classes). The weight of the ReCent loss is set to $1e-3$ ($\lambda$ of Equation (3.1)). Cosine similarity between the reconstruction centers is used as the margin parameter for the Derived-Margin softmax loss. Stochastic Gradient Descent optimizer [14] has been used for training the model with a batch size of 64, with an initial learning rate of 0.01, which reduces by a factor of 10 after every tenth epoch. The DeriveNet model is a classification model, trained for $n$ identities. During testing, the input VLR/LR image is passed through the feature extractor and the classification module, and the class with the highest score is chosen as the predicted label. For unseen testing, Cosine distance is used on the extracted embeddings. DeriveNet has been implemented in Python using the PyTorch framework [114], with a NVIDIA Tesla P100 GPU.

## 3.4 Datasets and Experimental Protocol

The DeriveNet model has been evaluated on three different tasks: (i) VLR/LR face recognition, (ii) VLR digit classification, and (iii) VLR/LR face recognition on drone-captured images. Details for each task are as below:

**(i) LR Face Recognition on the SCface dataset [42]:** The Surveillance Cameras Face (SCface) dataset contains face images of 130 subjects, captured from three different surveillance cameras in the visible range. Images are captured at a distance of 4.20, 2.60, and 1.00 meters. As per the existing protocol [81], images taken from 1.00m form the gallery, while images taken from $2.60 - 4.20$m act as the probes. 80 subjects form the training partition, and the remaining form the test set. Matching is performed at $128 \times 128$ resolution.

**(ii) VLR Face Recognition on the UCCS dataset [133]:** The UnConstrained College Students (UCCS) dataset [133] contains face images captured in a simulated surveillance environment. Images have been captured using a long-range surveillance camera kept at a stand-off distance of

around $100 - 150$ meters, without the knowledge of the subjects. The dataset contains labeled images of 1732 identities. As per the existing protocol [37, 138], images pertaining to 180 identities have been chosen (top 180 based on the number of images of each subject). The images are divided into training and testing partitions in a $4 : 1$ ratio. The testing set is down-sampled to $16 \times 16$ resolution, while the HR images are of $80 \times 80$. Consistent with the existing results, the rank-1 identification accuracy has been reported.

**(iii) VLR Digit Classification on the SVHN dataset [104]:** VLR digit classification has widespread applicability in surveillance scenarios such as license plate recognition and other digit classification applications. The Street View House Numbers (SVHN) dataset has been used for evaluating the DeriveNet model for VLR digit classification. The SVHN dataset contains images of $0 - 9$ digits captured from real-world natural scenes, having a resolution of $32 \times 32$. The dataset contains 73,257 images for training and 26,032 images for testing. The existing protocol [170] is followed, where the testing dataset is created by bicubically down-sampling the images by a factor of 4, to create images of $8 \times 8$ resolution, while the HR images are of $32 \times 32$ resolution. Consistent with existing results, the rank-1 (or top-1) and rank-5 (or top-5) accuracy has been reported.

**(iv) VLR/LR Face Recognition on the DroneSURF dataset [64]:** Face recognition in drone-shot videos has applicability in scenarios such as identifying individuals stuck at remote locations or in crowded places monitored via a drone. Recently, IARPAâĂŹs Biometric Recognition and Identification at Altitude and Range (BRIAR) program[1] also lays emphasis on the challenging problem of identifying individuals from long-range at elevated platforms. The DroneSURF dataset contains over 200 videos of 58 subjects captured across 411K frames. Pre-defined protocol is provided with the dataset, where 34 subjects form the training partition, and the remaining are used for testing. Along with the drone-shot videos, the dataset also contains four HR face images of each subject as the gallery images. Two protocols have been provided: (i) Active surveillance: where the drone actively follows the subjects, and (ii) Passive surveillance: where the drone monitors a particular area. The DeriveNet model has been evaluated on both the protocols (total 159,246 frames), and results have been reported in terms of the Cumulative Match Characteristic (CMC) curves (rank 1-5). As with the existing results [64], performance has been reported on the annotated face images provided with the dataset (most of them being smaller than $64 \times 64$). The HR gallery

[1] https://tinyurl.com/y9v7jkfv

Table 3.1: Rank-1 identification accuracy (%) on the SCface dataset [42] for LR face recognition. The table presents the performance of the DeriveNet model and the comparative accuracies on the protocol reported by Li *et al.* [81].

| **Algorithm** | **2.6m** | **4m** | **Avg.** |
|---|---|---|---|
| Coupled Mapping Method (2015) [135] | 43.24 | - | - |
| LMCM [190] | 60.40 | - | - |
| LMSoftmax for VLR (2016) [83] | 40.40 | - | - |
| L2Softmax for VLR (2017) [122] | 42.80 | - | - |
| AMSoftmax for VLR (2018) [164] | 46.80 | - | - |
| Large Margin Cosine Loss (2018) [165] | 53.12 | - | - |
| Angular Additive Margin Loss (2019) [20] | 57.58 | - | - |
| Centerloss for VLR (2019) [81] | 69.90 | - | - |
| **Student Teacher (2020) [95]** | **93.7** | 70.2 | 81.95 |
| **Proposed DeriveNet** | 92.80 | **76.80** | **84.80** |

images are of $128 \times 128$ resolution.

Figure 3-2 presents sample images from the datasets used for different applications. The HR and LR images are resized to the HR image dimension. Bicubic interpolation is used for conversion from HR to LR/VLR and vice-versa. For the face datasets (SCface and DroneSURF), facial regions are extracted using the coordinates given with each database, while the UCCS dataset provides cropped facial regions which are directly used. The SVHN database is used as is, without any additional cropping. For experiments on face recognition, the test set of the VGGFace2 dataset [15] (500 identities, 139K images) was used for pre-training the model with multi-resolution pyramid based data augmentation, while keeping all other implementation details consistent.

## 3.5 Results and Analysis

Tables 3.1-3.3 and Figures 3-5-3-9 present the performance and analysis of the DeriveNet model, along with other comparative techniques for all three case-studies. Each dataset's training and testing sets, as mentioned in Section 3.4, have been used for the analysis. Since existing protocols have been followed for all the datasets, comparative performance has directly been taken from the respective papers. The following paragraphs present case study wise analysis and the ablation study on DeriveNet:

(i) DeriveNet

(ii) Softmax based Model

Figure 3-5: Class Activation Maps obtained using the (i) DeriveNet model and the native (ii) Softmax based model. DeriveNet appears to focus more on the biometric regions of eyes and nose, while the Softmax based model appears to focus more on the hair and other soft features.

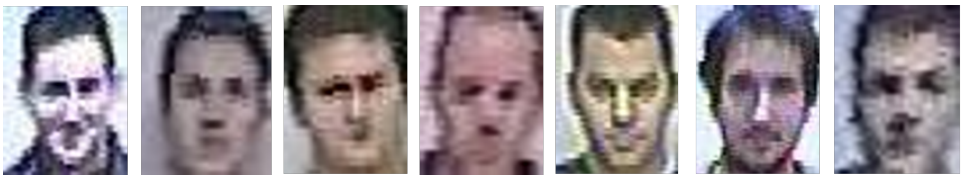


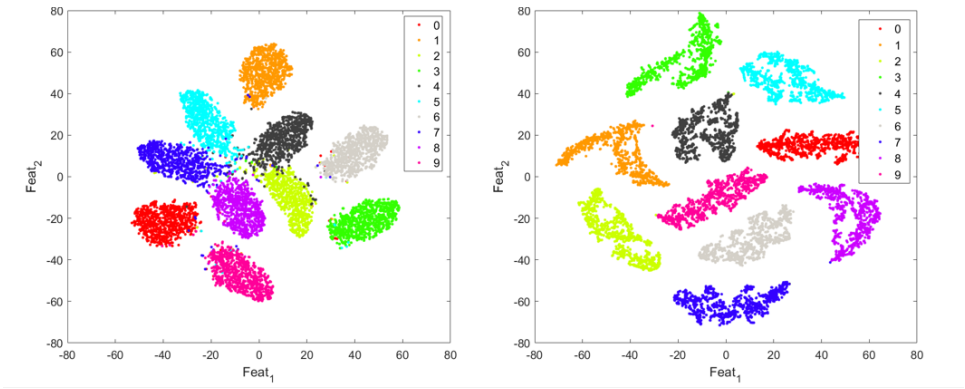
Figure 3-6: Sample images correctly identified via the proposed DeriveNet model, which are not identified by the other variants used for the ablation study.

**Analysis and Ablation Study for LR Face Recognition on the SCface Dataset:** Table 3.1 presents the rank-1 (top-1) recognition accuracy on the SCface dataset. The proposed model obtains a rank-1 identification accuracy of 92.80% and 76.80%, for $2.6m$ and $4m$, respectively, thus resulting in the current top-2 algorithms. An improvement of over 6% is observed for the farther distance ($4.6m$) as compared to the state-of-the-art algorithm [95]. Comparison has also been performed with other variants of the softmax loss, namely, Large-Margin softmax (LMSoftmax), Angular-Margin softmax (AMSoftmax), and $\ell_2$-constrained softmax (L2Softmax). Owing to the same protocol, comparative results have directly been taken from the published manuscript [81]. Comparison has also been performed with the Large Margin Cosine loss [165] and the Angular Additive Margin loss [20]. The DeriveNet model presents an improvement of over 20% from the above mentioned existing softmax-based loss functions for $2.6m$ distance. We believe that the incorporation of HR information during training, and the explicit modeling of the inter-class variations promotes learning of an improved classifier. We also performed the Pearson's Chi-square Test for independence between the top two results and a statistical significance was obtained with a confidence interval of 90% (p-value$<$0.01).

In order to understand the effect of each component of the DeriveNet model, an ablation study

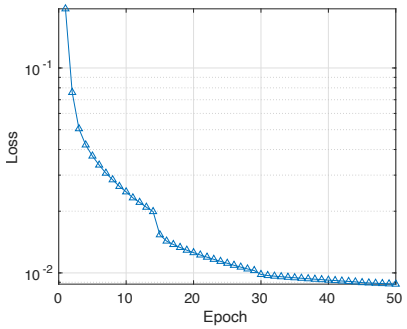
is performed. As demonstrated from Table 3.1 an average accuracy of 84.80% is observed by the DeriveNet model on the SCface dataset, across both the distances. Upon removing the component of multi-resolution pyramid augmentation for pre-training, the DeriveNet model reports an accuracy of 75.40%, thus demonstrating the effectiveness of the augmentation based pre-training. Further, in order to analyze different components of the DeriveNet model only, performance has been reported for two other variants of the model which also do not utilize the multi-resolution pyramid augmentation for pre-training: 'DeriveNet - {D-Margin}',where the Derived-Margin softmax loss is replaced with a traditional softmax loss, and 'DeriveNet - {ReCent}', where the ReCent loss is not applied. An accuracy of 71.20% is obtained without the Derived-Margin softmax loss, demonstrating a drop of around 4.2% from the DeriveNet model. On removing the ReCent loss (a fixed margin of 0.2 is used by the proposed Derived-Margin softmax loss), an accuracy of 72.54% is obtained. As compared to the earlier combination, a smaller drop in accuracy is observed, suggesting a higher contribution of the proposed Derived-Margin softmax loss in the DeriveNet model. Overall, a drop of around $2.9 - 4.2\%$ is observed upon removing any component of the DeriveNet model, thus encouraging the inclusion of each component. We also performed the McNemar test [96] between the predicted labels of DeriveNet and other variants used in the ablation study and achieved statistical difference at a confidence interval of 99% ($p$-value$<$0.01). Figure 3-5 presents the Class Activation Maps [196] obtained via the DeriveNet model and the Softmax based model. The DeriveNet model appears to focus on key biometric features such as the eyes and the nose, whereas the Softmax based model focuses more on the chin and hair region. Figure 3-6 presents sample images mis-classified by the variants of the DeriveNet model used for the ablation study, correctly classified by DeriveNet, thus demonstrating its effectiveness.

In some real world cases, it might be difficult to obtain corresponding LR-HR pairs for training. In order to simulate such scenarios, we also evaluated the the DeriveNet model with non-mated LR-HR pairs during training. In this case, each LR image was paired with a random HR image of the same class. The DeriveNet model achieves a recognition performance of 75.00% when trained with this data, demonstrating a reduction of only 0.44% as compared to training with mated pairs (without multi-resolution data augmentation based pre-training), thus promoting the usability of the DeriveNet model even with unavailability of mated pairs during training. Experiments were also performed to reinstate the benefit of deriving the margins from the reconstructed HR space

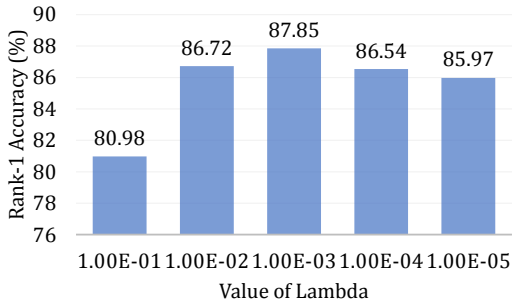



(i) Softmax based Model (ii) DeriveNet

(a) t-SNE Visualizations




(b) Loss Plot




(c) Effect of $\lambda$

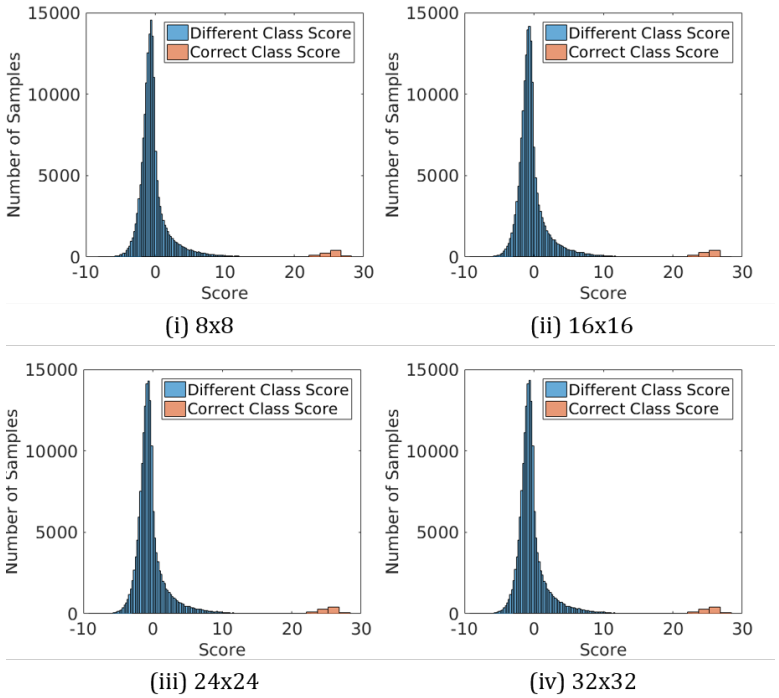
Figure 3-7: (a) t-SNE plots on the features of (i) Softmax based model and (ii) DeriveNet model. A larger margin is observed between the features learned via DeriveNet. (b) Convergence plot of the DeriveNet model, and (c) Effect of $\lambda$ (Equation (3.1)) on the SVHN dataset.

as opposed to the original HR space. An accuracy of 73.43% is obtained when margins are derived from the original HR space, demonstrating a drop of 2% from the proposed DeriveNet model (without multi-resolution data augmentation based pre-training), thus supporting using the reconstructed HR space for margins.

**Analysis for VLR Face and Digit Recognition on the UCCS and SVHN Datasets:** Table 3.2 presents the rank-1 (top-1) recognition performance on the UCCS face dataset. As compared to the state-of-the-art result, the DeriveNet model demonstrates an improvement of around 1.7%, by obtaining an accuracy of 97.57%. Comparison has also been performed with some of the softmax-based losses, where, an improvement of at least 12% is observed by the DeriveNet model. Owing to the same protocol, the accuracies have directly been reported from the published paper

Table 3.2: Rank-1 accuracy (%) on the UCCS dataset [133] for VLR face recognition ($16 \times 16$), with $80 \times 80$ HR images.

| **Algorithm** | **Acc. (%)** |
|---|---|
| Robust Partially Coupled Nets (2016) [170] | 59.03 |
| LMSoftmax for VLR (2016) [83] | 64.90 |
| L2Softmax for VLR (2017) [122] | 85.00 |
| AMSoftmax for VLR (2018) [164] | 58.60 |
| SICNN (2018) [188] | 93.38 |
| Selective Knowledge Distillation (2019) [37] | 67.25 |
| Coupled GAN (2019) [158] | 71.68 |
| Centerloss for VLR (2019) [81] | 93.40 |
| **DirectCapsNet (2019) [138]** | **95.81** |
| CSRIP (2020) [43] | 93.49 |
| Bridge Distillation (2020) [38] | 81.92 |
| Hybrid Order Distillation (2020) [36] | 77.81 |
| **Proposed DeriveNet** | **97.57** |



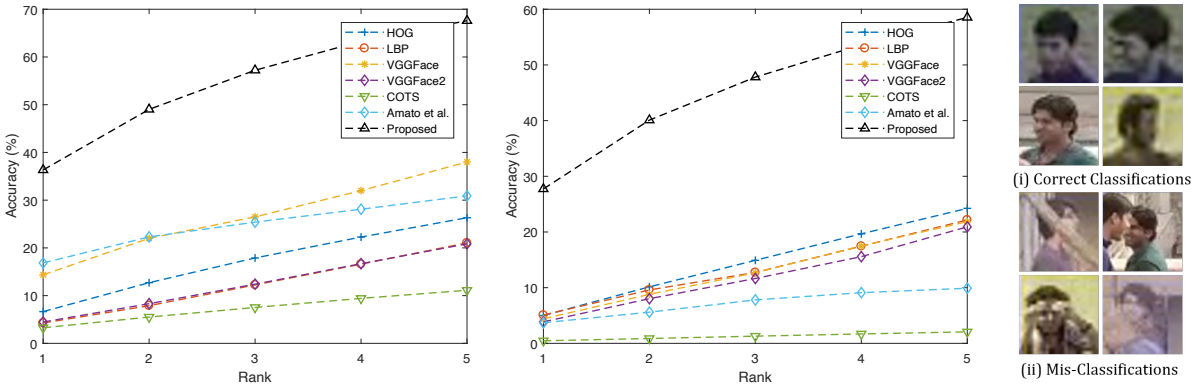

(i) 8x8 (ii) 16x16

(iii) 24x24 (iv) 32x32

Figure 3-8: Score distributions of the correct and incorrect classes for different resolutions on the UCCS dataset. The Earth Mover's Distance was also calculated between the same class and different class scores for each resolution: 24.55 ($8 \times 8$), 24.65 ($16 \times 16$), 24.72 ($24 \times 24$), and 24.71 ($32 \times 32$). The distance for the smallest resolution ($8 \times 8$) is the least, thus suggesting a larger variation between the scores, as compared to the other resolutions.

Table 3.3: Top-1 and top-5 accuracy (%) on the SVHN dataset [104] for VLR digit recognition ($8 \times 8$).

| **Algorithm** | **Accuracy (%)** | |
|---|---|---|
| | Top-1 | Top-5 |
| CNN (VLR) (2016) [170] | 45.29 | 66.78 |
| RPC Nets (2016) [170] | 56.98 | 70.82 |
| SICNN (2018) [188] | 81.53 | 96.77 |
| CapsNet (VLR) (2019) [138] | 79.19 | 88.89 |
| DirectCapsNet (2019) [138] | 84.51 | 91.20 |
| **CSRIP (2020) [43]** | **84.62** | **97.32** |
| **Proposed DeriveNet** | **87.85** | **97.18** |




(a) Active Surveillance (b) Passive Surveillance (c) Samples

Figure 3-9: (a-b) CMC curves on the DroneSURF dataset for the two protocols. DeriveNet improves the rank-1 accuracy by over 20% on both protocols. Comparison has also been performed with Amato *et al.* [5]. (c) Sample challenging cases from the DroneSURF dataset which were (i) correctly classified and (ii) mis-classified by the proposed DeriveNet model.

[81]. Further, comparison has also been performed with recent super-resolution based algorithms [43, 188], where an improved performance is obtained via the DeriveNet model. The model has also been evaluated with different input resolutions, specifically, $8 \times 8$, $24 \times 24$, and $32 \times 32$, where the proposed model achieves an accuracy of 97.35%, 97.79%, and 97.79%, respectively. Figure 3-8 presents the score distribution of the correct and incorrect classes, demonstrating large difference between the two at the varying resolutions.

Table 3.3 presents the rank-1 (top-1) and rank-5 (top-5) classification accuracies on the SVHN dataset. The DeriveNet model obtains a rank-1 classification accuracy of 87.85%, demonstrating an improvement of over 3% from the state-of-the-art result. We also analyzed the value of the

margin parameter for the Derived-Margin softmax loss: a margin of 0.38 is obtained for the digits 6 and 7, while a larger margin of 0.66 is observed for the digits 5 and 6. Since the margin parameter encodes the class similarity, a larger margin demonstrates a larger similarity between the classes 5 and 6, which thus enforces a larger score difference between the two classes. Figure 3-7(a) presents the t-SNE visualizations [74] of the features learned by the two models. The DeriveNet model appears to learn features with larger inter-class margins, which we believe result in better classification at low resolutions. Figure 3-7(b) presents the training loss curve of the DeriveNet model, demonstrating the converging nature of the training process. Further, Figure 3-7(c) presents the accuracy of the proposed model with different values of $\lambda$ (Equation (3.1)) signifying the weight given to the ReCent loss. A higher value results in reduced performance since the model focuses more on the reconstruction space as compared to learning a discriminative classifier (Derived-Margin softmax loss). The improved performance demonstrates that the DeriveNet model can also be applied for object classification (e.g., digits) as well.

For both the datasets (UCCS and SVHN), the Pearson's Chi-square Test for independence was performed between the top two results and a statistical significance was obtained with a confidence interval of 95% ($p-$value$<$0.05).

**Analysis for VLR/LR Face Recognition on the DroneSURF Dataset:** Figure 3-9 presents the CMC curves on the challenging DroneSURF dataset. Comparison has also been performed with the results reported by Amato *et al.* [5]. For the active surveillance protocol (Figure 3-9(a)), where the drone actively follows an individual, an improvement of around 20% is observed at rank-1 by the DeriveNet model, where it achieves 36.33%. Similar results are obtained on the passive surveillance protocol (Figure 3-9(b)), where the proposed model performs better than the existing results at all the ranks (rank 1-5). At rank-5, the proposed model achieves a recognition performance of 58.53%, demonstrating an improvement of over 33% from the current state-of-the-art. Figure 3-9(c) presents sample face images that are classified correctly and incorrectly by the proposed model. Despite the relatively lower resolution of face images, the DeriveNet model is able to recognize faces under varying pose and expression as well (Figure 3-9(c)(i)). However, cases with severe occlusion resulting in partial faces still remain as some of the key challenges for drone-based face recognition (Figure 3-9(c)(ii)). The improved performance of DeriveNet on the

Table 3.4: Performance of the DeriveNet model with and without the Multi-resolution Pyramid based data augmentation during training. The tabulated metric (%) varies across datasets and is consistent with the ones reported earlier in the manuscript: SCface (avg. rank-1 accuracy), UCCS (rank-1 accuracy), SVHN (rank-1 accuracy), DroneSURF (rank-5 accuracy), and QMUL-SurvFace (avg. accuracy).

| **Dataset** | **Pyramid Augmentation** | |
|---|---|---|
| | **Without** | **With** |
| SCface [42] | 75.44 | 84.80 |
| UCCS [133] | 97.57 | 97.57 |
| SVHN [104] | 87.85 | 87.85 |
| DroneSURF (Active) [64] | 65.79 | 67.60 |
| DroneSURF (Passive) [64] | 48.56 | 58.53 |
| QMULSurvFace [18] | 68.37 | 72.34 |

DroneSURF dataset promotes its applicability for LR drone-based face recognition, and its utility in critical scenarios.

**Additional Experiment on the QMUL-SurvFace Dataset (Verification Protocol) [18]:** Experiments have been performed on the pre-defined verification protocol [18]. DeriveNet achieves a mean accuracy of 68.37%, while the current state-of-the-art algorithm, Student-Teacher model [95] achieves 72%. The paper also provides results on super-resolved images (61%) and the base model used by the authors (56%). With the multi-resolution pyramid data augmentation based pre-training, the DeriveNet model achieves an accuracy of 72.34%. The comparative performance of the DeriveNet model further promotes its usage for such challenging unseen scenarios as well.

**Impact of Multi-Resolution Pyramid based Data Augmentation:** Table 3.4 presents the performance of the DeriveNet model on different datasets with and without the Multi-resolution Pyramid based data augmentation. An improvement of at most around 10% is observed by incorporating the multi-resolution pyramid based data augmentation during model training. Maximum improvement is observed for datasets and settings where during evaluation the trained model is provided unconstrained facial images having varying resolution (similar to real-world settings). It is our belief that since the pyramid augmentation enables the model to capture the variations across different resolutions during training, it results in boosted performance. On the other hand, less (or no) improvement is observed in setups where evaluation is performed on images captured in relatively constrained setups or having lesser resolution variations.

## 3.6 Summary

Classification of (very) low resolution images has wide-spread applicability in various scenarios such as image tagging, person identification in surveillance settings, and vehicle number-plate recognition. VLR/LR image regions often suffer from the challenge of limited interpretive information content, making it difficult to extract discriminative representations. This research proposes a novel DeriveNet model for VLR/LR image classification, which consists of two loss functions: (i) proposed Derived-Margin softmax loss, and the (ii) Reconstruction-Center loss, termed as the ReCent loss. The proposed Derived-Margin softmax loss incorporates a margin/penalty into the native softmax loss for modeling the inter-class variations. The DeriveNet model has been trained with a novel multi-resolution pyramid based data augmentation technique for modeling the resolution variations. Experimental results demonstrate the superiority of the DeriveNet model, even in challenging scenarios such as VLR/LR face recognition from drone-shot images (over 15% improvement at rank-1 for different protocols). The ablation study supports inclusion of each component in the DeriveNet model (drop of $2.9 - 4.2\%$ is observed upon removal of any component). Further, superior performance is obtained as compared to several recent algorithms on different tasks and datsets (including digit classification and face recognition in drone-shot images). This research has been published in the IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE T-PAMI), and all the figures have been taken from the published manuscript [139].

THIS PAGE INTENTIONALLY LEFT BLANK

# Chapter 4

# Disguised Faces in the Wild

## 4.1 Introduction

Extensive research in the domain of face recognition has resulted in the development of algorithms achieving state-of-the-art performance on large-scale unconstrained datasets [28, 41, 86, 94]. However, it has often been observed that most of these systems are susceptible to digital and physical adversaries [4, 33, 40, 62, 145, 179]. Digital adversaries refer to manipulations performed on the image being provided to the recognition system, with the intent of *fooling* the system. It has been shown that traditional systems based on hand crafted features [40] degrade gracefully with digital attacks while deep learning systems deteriorate rapidly. Recently, the issue of digital attacks has garnered attention, with perturbation techniques such as Universal Adversarial Perturbation [100] and DeepFool [101] demonstrating devastating adversarial performance on different algorithms. On the other hand, physical adversaries refer to the variations brought to the individual before capturing the input data for the recognition system. In case of face recognition, this can be observed due to variations caused by different spoofing techniques or disguises. While the area of spoof detection and mitigation is being well explored [33, 120], research in the domain of disguised face recognition is yet to receive dedicated attention, despite its significant impact on both traditional and deep learning systems [26, 73].

Disguised face recognition encompasses handling both *intentional* and *unintentional* disguises. Intentional disguise refers to the scenario where a person attempts to hide his/her identity or *impersonate* another person's identity, in order to fool a recognition system into obtaining unauthorized

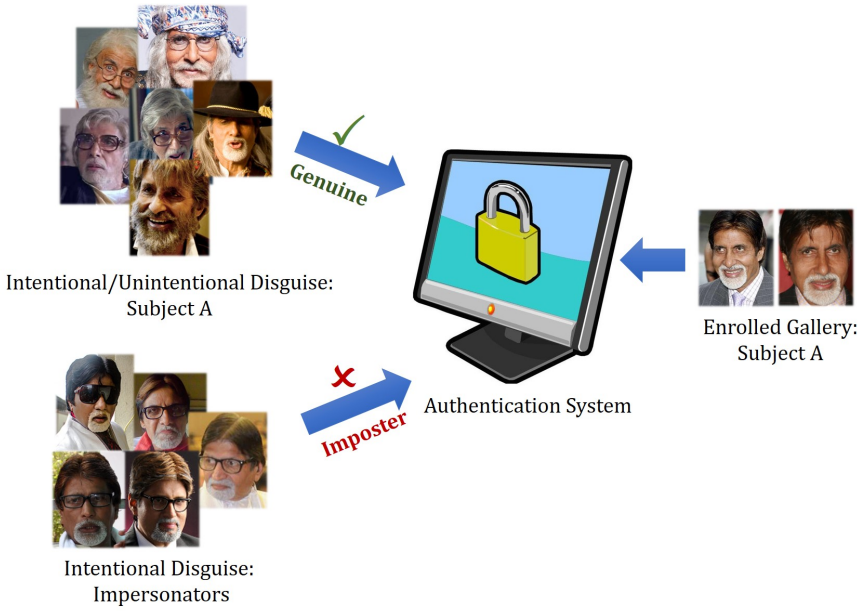



Figure 4-1: Authentication systems often face the challenge of matching disguised face images with non-disguised enrolled images. Figure has been taken from the published manuscript [143].

access. This often results in utilizing external disguise accessories such as wigs, beard, hats, mustache, and heavy makeup, leading to obfuscation of the face region. This renders low inter-class variations between different subjects, thereby making the problem challenging in nature. Unintentional disguises cover a range of images wherein the face is obfuscated by means of an accessory such as glasses, hats, and masks. It can also be due to aging, resulting in an increase or decrease of facial hair such as beard or mustache, and variations in the skin texture. Unintentional disguises create challenges for the face recognition system by increasing the intra-class variations for a given subject. The combination of both intentional and unintentional disguises render the problem of disguised face recognition an arduous task. Figure 4-1 presents sample images of intentional and unintentional disguises, along with non-disguised enrolled face images. The authentication system faces the challenge of verifying an image containing unconstrained disguise variations against a frontal non-disguised face image.

This chapter presents the Disguised Faces in the Wild 2018 (referred to as the DFW or DFW2018) dataset and the Disguised Faces in the Wild 2019 (referred to as the DFW2019) dataset. The DFW2018 dataset contains 11,157 face images of 1,000 identities. Almost the entire dataset is collected from the Internet resulting in an unconstrained set of images. One of the key highlights

of the dataset is the availability of (i) *normal*, (ii) *validation*, (iii) *disguised*, and (iv) *impersonator* images for a given subject. This is a unique dataset containing multiple types of in-the-wild images for a subject in order to evaluate different aspects of disguised face recognition, along with three pre-defined evaluation protocols. Here, for a given subject, disguised face images are images of the same subject with disguise accessories, while impersonators correspond to images of different subjects. It is our assertion that the availability of a large-scale dataset, containing images captured in unconstrained settings across multiple devices, pose, illumination, and disguise accessories would help in encouraging research in this direction. The dataset was released as part of the DFW challenge, in the Disguised Faces in the Wild Workshop at International Conference on Computer Vision and Pattern Recognition (CVPR), 2018. The Disguised Faces in the Wild 2019 (DFW2019) dataset builds upon the DFW2018 dataset and contains 3840 face images of 600 subjects, having variations across disguise accessories, bridal make-up, and plastic surgery. The dataset contains four pre-defined protocols for face verification, along with the availability of images with bridal make-up and plastic surgery. The dataset was released in the DFW2019 competition, as part of the International Workshop on Disguised Faces in the Wild, held in conjunction with the International Conference on Computer Vision (ICCV), 2019. We present the DFW2018 and DFW2019 datasets, along with the findings across the different evaluation protocols. Performance of participants in the DFW challenges, baseline results, and analysis of different difficulty levels have also been provided.

## 4.2 Motivation

Table 4.1 presents the characteristics of existing disguise face datasets, along with the DFW2018 and DFW2019 datasets. One of the initial datasets containing disguise variations is the AR dataset [91]. It was released in 1998 and contains a total of 3,200 face images having some images containing controlled disguise variations. This was followed by the release of different datasets having variations across disguise accessories and dataset size. Most of the datasets are moderately sized having controlled disguise variations. Other than disguised face datasets, a lot of recent research in face recognition has focused on large-scale datasets captured in unconstrained environments [15, 57, 67, 103, 175]. The availability of such datasets facilitate research in real world scenarios,

Table 4.1: Summary of disguise face datasets in literature.

| **Name** | **Controlled Disguise** | **Number of** | | **Availability of Impersonators** | **Publicly Available** |
|---|---|---|---|---|---|
| | | **Images** | **Subjects** | | |
| AR Dataset (1998) [91] | Yes | 3,200 | 126 | No | Yes |
| National Geographic Dataset (2004) [121] | Yes | 46 | 1 | No | No |
| Synthetic Disguise Dataset (2009) [145] | Yes | 4,000 | 100 | No | No |
| Curtin Faces Dataset (2011) [79] | Yes | 5,000 | 52 | No | Yes |
| IIITD $I^2$BVSD Dataset (2014) [24] | Yes | 1,362 | 75 | No | Yes |
| Disguised and Makeup Faces Dataset (2016) [167] | No | 2,460 | 410 | No | Yes |
| Spectral Disguise Face Dataset (2018) [119] | Yes | 6,480 | 54 | No | Yes |
| **DFW Dataset (2018)** | **No** | **11,157** | **1,000** | **Yes** | **Yes** |
| **DFW2019 Dataset (2019)** | **No** | **3,840** | **600** | **Yes** | **Yes** |

however, they do not focus on the aspect of disguised face recognition.

Disguised face recognition presents the challenge of matching faces under both intentional and unintentional distortions. It is interesting to note that both forms of disguise can result in either genuine or imposter pairs. For instance, a criminal may intentionally attempt to conceal his identity by using external disguise accessories, thereby resulting in a genuine match for an authentication system. On the other hand, an individual might intentionally attempt to impersonate another person, resulting in an imposter pair for the face recognition system. Similarly, in case of unintentional disguises, use of casual accessories such as sunglasses or hats results in a genuine disguised pair, while individuals who look alike are imposter pairs for the recognition system. The combination of different disguise forms along with the intent makes the given problem more challenging.

To the best of our knowledge, no existing disguise dataset captures the wide spectrum of intentional and unintentional disguises. To this effect, we prepared and released the DFW2018 and DFW2019 datasets. The datasets simulate the real world scenario of unconstrained disguise variations, and provides multiple impersonator images for almost all subjects. The presence of impersonator face images enables the research community to analyze the performance of face recognition models under physical adversaries. The datasets were released as part of the DFW workshops, where researchers from all over the world were encouraged to evaluate their algorithms against this challenging task. For the DFW2018 dataset, inspired by the presence of disguise intent in real world scenarios, algorithms were evaluated on three protocols: (i) Impersonation, (ii) Obfuscation, and (iii) Overall. For the DFW2019 dataset, algorithms were also evaluated on an additional plastic surgery protocol. Impersonation focuses on disguise variations where an individual either attempts to impersonate another individual intentionally, or looks like another individual unintentionally. In both cases, the authentication system should be able to detect an (imposter) unauthorized access attempt. The second protocol, obfuscation, focuses on intentional or unintentional disguise variations across genuine users. In this case, the authentication system should be able to correctly identify genuine users even under varying disguises. The third protocol evaluates a face recognition model on the entire dataset. As mentioned previously, it is our hope that the availability of the DFW2018 and DFW2019 datasets along with the pre-defined protocols would enable researchers to develop state-of-the-art algorithms robust to different physical adversaries.

## 4.3 Disguised Faces in the Wild (DFW) 2018 Dataset

As shown in Table 4.1, most of the research in the field of disguised face recognition has focused on images captured in controlled settings, with limited set of accessories. In real world scenarios, the problem of disguised face recognition extends to data captured in uncontrolled settings, with large variations across disguise accessories. Combined with the factor of *disguise intent*, the problem of disguise face recognition is often viewed as an exigent task. The DFW dataset simulates the above challenges by containing 11,157 face images belonging to 1,000 identities with uncontrolled disguise variations. It is the first dataset which also provides impersonator images for a given subject. The DFW dataset contains the IIIT-Delhi Disguise Version 1 Face Database (ID V1) [26] having 75 subjects, and images corresponding to the remaining 925 subjects have been taken from the Internet. Since the images have been taken from the Web, most of the images correspond to famous personalities and encompass a wide range of disguise variations. The dataset contains images with respect to unconstrained disguise accessories such as hair-bands, masks, glasses, sunglasses, caps, hats, veils, turbans, and also variations with respect to hairstyles, mustache, beard, and make-up. Along with the disguise variations, the images also demonstrate variations across illumination, pose, expression, background, age, gender, and camera quality. The dataset is publicly available for research purposes and can be downloaded from our website [1]. The following subsections present the dataset statistics, protocols for evaluation, and details regarding data distribution.

### 4.3.1 Dataset Statistics

As mentioned previously, the DFW dataset contains images pertaining to 1,000 identities, primarily collected from the Internet. Most of the subjects are adult famous personalities of Caucasian or Indian ethnicity. Each subject contains at least five and at most twenty six images. The dataset comprises of 11,157 face images including different kinds of images for a given subject, that is, *normal*, *validation*, *disguised*, and *impersonator*. Detailed description of each type is given below:

- **Normal Face Image:** Each subject has a frontal, non-disguised, good quality face image, termed as the normal face image.

[1] http://iab-rubric.org/resources/dfw.html


(a) Subject A


(b) Subject B

Figure 4-2: Images pertaining to two subjects of the DFW dataset. The dataset contains at most four types of images for each subject: Normal, Validation, Disguised, and Impersonator.

- **Validation Face Image:** Other than the normal face image, 903 subjects have another non-disguised face image, referred to as the validation image. This can help in evaluating a proposed model for matching non-disguised face images.

- **Disguised Face Image:** For each subject, disguised face images refer to images having intentional or unintentional disguise of the same subject. For the 1,000 identities present in the dataset, every subject has at least one and at most 12 disguised images. These images form genuine pairs with the normal and validation face images, and can help in evaluating the true positive rate of an algorithm.

- **Impersonator Face Image:** Impersonators refer to people who intentionally or unintentionally look similar to another person. For a given subject, impersonator face images belong to different people, thereby resulting in imposter pairs which can be used to evaluate the true negative rate of an algorithm. The images were collected from the Internet using different relevant keywords on Google Images, news articles, and popular entertainment blogs and later manually verified by human examiners. In the DFW dataset, 874 subjects have images

Table 4.2: Statistics of the DFW dataset.

| Characteristic | Count |
|---|---|
| Subjects | 1,000 |
| Images | 11,157 |
| Normal Images | 1,000 |
| Validation Images | 903 |
| Impersonator Images | 4,440 |
| Range of Images per Subject | [5,26] |

Table 4.3: Statistics of the training and testing sets of the DFW dataset.

| Number of | Training Set | Testing Set |
|---|---|---|
| Subjects | 400 | 600 |
| Images | 3,386 | 7,771 |
| Normal Images | 400 | 600 |
| Validation Images | 308 | 595 |
| Disguised Images | 1,756 | 3,058 |
| Impersonator Images | 922 | 3,518 |

corresponding to their impersonators, each having at least 1 and at most 21 images.

Statistics of the DFW dataset are presented in Table 4.2, and Figure 4-2 demonstrates sample images of two subjects. It can be observed that disguised face images result in increased intra-class variations for a given subject, while the impersonator images render lower inter-class variability. Overall, the DFW dataset contains 1,000 and 903 normal and validation face images, respectively, along with 4,814 disguised face images, and 4,440 impersonator images.

### 4.3.2 Protocols for Evaluation

The DFW dataset has been released with three protocols for evaluation. A fixed training and testing split is provided which ensures mutual exclusion of images and subjects. Images from four hundred subjects are used to create the training set, and the remaining six hundred subjects form the test set. Table 4.3 presents the statistics of the testing and training sets. All three protocols correspond to verification, where a face recognition module is expected to classify a pair of images as genuine or imposter. Detailed description of each protocol on the pre-defined training and testing partitions is given below:

**Protocol-1 (Impersonation)** evaluates a face recognition model for its ability to distinguish impersonators from genuine users with high precision. A combination of a normal image with a validation image of the same subject corresponds to a genuine pair for this protocol. For imposter pairs, the impersonator images of a subject are partnered with the normal, validation, and disguised images of the same subject.

**Protocol-2 (Obfuscation)** is useful for evaluating the performance of a face recognition system under intentional or unintentional disguises, wherein a person attempts to hide his/her identity. The genuine set contains pairs corresponding to the (normal, disguise), (validation, disguise), and (disguise$_1$, disguise$_2$) images of a subject. Here, disguise$_n$ corresponds to the $n^{th}$ disguised image of a subject. That is, all pairs generated using the normal and validation images with the disguise images, and the pairs generated between the disguise images of the same subject, constitute the genuine pairs. The imposter set is created by combining the normal, validation, and disguised images of one subject with the normal, validation, and disguised images of a different subject. This results in the generation of cross-subject imposter pairs. The impersonator images are not used in this protocol.

**Protocol-3 (Overall Performance)** is used to evaluate the performance of a given face recognition algorithm on the entire DFW dataset. The genuine and imposter sets created in the above two protocols are combined to generate the data for this protocol. For the genuine set, pairs are created using the (normal, validation), (normal, disguise), (validation, disguise), and (disguise$_1$, disguise$_2$) images of the same subject. For the imposter set, cross-subject imposter pairs are considered, wherein the normal, validation, and disguised face images of one subject are combined with normal, validation, and disguised face images of another subject. Apart from the cross-subject imposter pairs, the impersonators of one subject are also combined with normal, validation, and disguised face images of the same subject to further supplement the imposter set.

### 4.3.3 Nomenclature and Data Distribution

The DFW dataset is available for download as an archived file containing one folder for each subject. Each of the 1,000 folders is named with the subject's name and may contain the four types of images discussed above: normal, validation, disguise, and impersonator. In order to ensure

consistency and eliminate ambiguity, the following nomenclature has been followed across the dataset:

- Each subject has a single normal face image, which has been named as *firstName_lastName.jpg*. For instance, for the subject Alicia Keys, the subject's normal image is named *Alicia_Keys.jpg*.
- As mentioned previously, a given subject contains only a single validation face image. Therefore, the validation image is named with a postfix '_a', that is, *firstName_lastName_a.jpg*. For the example of Alicia Keys, the subject validation image is stored as *Alicia_Keys_a.jpg*.
- For disguised face images, a postfix of '_h' is adopted, along with a number for uniquely identifying the disguised face image of a given subject. That is, *firstName_lastName_h_number.jpg*. Here, *number* can take values such as '001', '002', ... '010'. For example, the first disguise image of subject Alicia Keys can be named as *Alicia_Keys_h_001.jpg*, while the third disguised face image can be named as *Alicia_Keys_h_003.jpg*.
- Similar to the disguised image nomenclature, a postfix of '_I' is used to store the impersonator images of a subject. That is, impersonator images are named as *firstName_lastName_I_number.jpg*. For example, the first impersonator image of subject Alicia Keys can be named as *Alicia_Keys_I_001.jpg*.

In order to correctly follow the protocols mentioned above, and report corresponding accuracies, training and testing mask matrices are also provided along with the dataset. Given the entire training or testing partition, the mask matrix can be used to extract relevant genuine and imposter pairs or scores for a given protocol. The DFW dataset also contains face co-ordinates obtained via faster RCNN [125]. Given an image of the dataset, the co-ordinates provide the face location in the entire image.

## 4.4 Disguised Faces in the Wild 2018 Competition

Disguised Faces in the Wild competition was conducted as part of the *First International Workshop on Disguised Faces in the Wild*[2], at the International Conference on Computer Vision and Pattern

[2]http://iab-rubric.org/DFW/dfw.html

Recognition, 2018 (CVPR'18). Participants were required to develop a disguised face recognition algorithm, which was evaluated on all three protocols of the DFW dataset. The competition was open world-wide, to both industry and academic institutions. The competition saw over 100 registrations from across the world.

All participating teams were provided with the DFW dataset, including the training and testing splits, face co-ordinates, and mask matrices for generating the genuine and imposter pairs. Evaluation was performed based on the three protocols described in Section 4.3.2. No restriction was enforced in terms of utilizing external training data, except ensuring mutual exclusion with the test set. The remainder of this section presents the technique and performance analysis of all the submissions, including the baseline results.

### 4.4.1 Baseline Results

Baseline results are computed using the VGG-Face descriptor [113], which is one of the top performing deep learning models for face recognition. A pre-trained VGG-Face model is used for feature extraction (trained on the VGG-Face dataset [113]). Baseline results were also provided to the participants. Baseline results have also been computed with the ResNet-50 architecture trained on the MS-Celeb-1M and VGGFace2 datasets [15]. The extracted features are compared using Cosine distance, followed by classification into genuine or imposter. Both the models achieve high recognition performance on challenging face datasets.

### 4.4.2 DFW2018 Competition: Submissions

The DFW competition received 12 submissions from all over the world, having both industry and academic affiliations. Table 4.4 presents the list of the participating teams, along with their affiliation. Details regarding the technique applied by each submission are provided below:

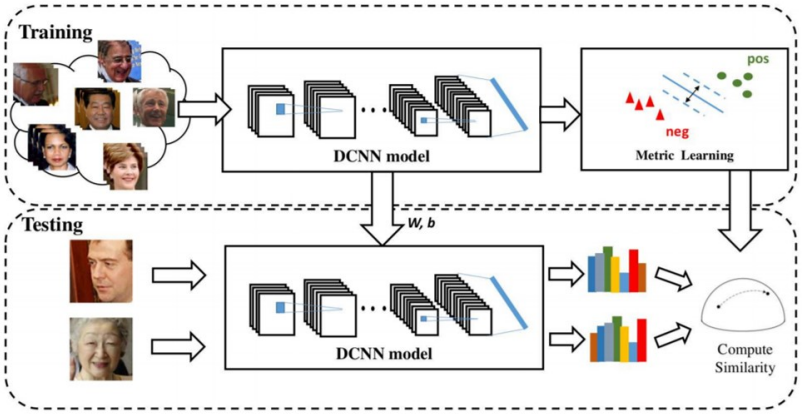
**(i) Appearance Embeddings for Face Representation Learning (AEFRL) [148]:** AEFRL is a submission from the Information Technologies, Mechanics and Optics (ITMO), Russian Federation. Later in the competition, it was renamed to Hard Example Mining with Auxiliary Embeddings. Faces are detected, aligned, and cropped using Multi-task Cascaded Convolutional Net-

Table 4.4: List of teams which participated in the DFW competition.

| **Model** | **Affiliation** | **Brief Description** |
|---|---|---|
| AEFRL [148] | ITMO University, Russia | MTCNN + 4 networks for feature extraction + Cosine distance |
| ByteFace | Bytedance Inc., China | Weighted linear combination of ensemble of 3 CNNs |
| DDRNET [70] | West Virginia University, USA | Inception Network with Center Loss |
| DisguiseNet [115] | Indian Institute of Technology Ropar, India | Siamese network with VGG-Face having a weighted loss |
| DR-GAN | Michigan State University, USA | MTCNN + DR-GAN + Cosine distance |
| LearnedSiamese | Computer Vision Center UAB, Spain | Cropped faces + Siamese Neural Network |
| MEDC | Northeastern University, USA | MTCNN + Ensemble of 3 CNNs + Average Cosine distance |
| MiRA-Face [187] | National Taiwan University, Taiwan | MTCNN + RSA + Ensemble of CNNs |
| OcclusionFace | Zhejiang University, China | MTCNN + Fine-tuned ResNet-28 |
| Tessellation | Tessellate Imaging, India | Siamese network with triplet loss model |
| UMDNets [7] | The University of Maryland, USA | All-In-One + Average across scores obtained by 2 networks |
| WVU_CVL | West Virginia University, USA | MTCNN + CNN + Softmax |

| Description | Output |
|---|---|
| input image | $255 \times 255 \times 3$ |
| $5 \times 5 \times 32$ Conv, stride 2 | $126 \times 126 \times 32$ |
| $3 \times 3 \times 64$ Conv | $124 \times 124 \times 64$ |
| $2 \times 2$ MaxPool, stride 2 | $62 \times 62 \times 64$ |
| $3 \times 3 \times 64$ ResBlock | $62 \times 62 \times 64$ |
| $3 \times 3 \times (96 + 64 + 32)$ Conv, d=1,2,3 | $60 \times 60 \times 192$ |
| $2 \times 2$ MaxPool, stride 2 | $30 \times 30 \times 192$ |
| $(3 \times 3 \times 192$ SE-ResBlock, r=16)$\times 2$ | $30 \times 30 \times 192$ |
| $3 \times 3 \times (336 + 112)$ Conv, d=1,2 | $28 \times 28 \times 448$ |
| $2 \times 2$ MaxPool, stride 2 | $14 \times 14 \times 448$ |
| $(3 \times 3 \times 448$ SE-ResBlock, r=16)$\times 5$ | $14 \times 14 \times 448$ |
| $3 \times 3 \times 1024$ Conv | $12 \times 12 \times 1024$ |
| $2 \times 2$ MaxPool, stride 2 | $6 \times 6 \times 1024$ |
| $(3 \times 3 \times 1024$ SE-ResBlock, r=16)$\times 5$ | $6 \times 6 \times 1024$ |
| $512 + 512$, fc + maxout, group = 2 | 512 |
| $L_2$-normalization | 512 |

(a) AEFRL




(b) UMDNets

Figure 4-3: Diagrammatic representation of (a) AEFRL [148], and (b) UMDNets [7]. Images have been taken from their respective publications.

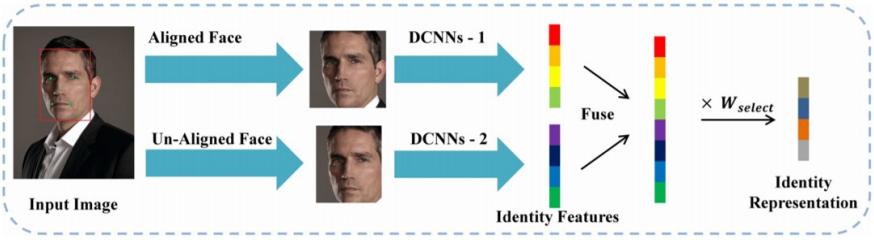



Figure 4-4: Diagrammatic representation of MiRA-Face [187]. Image has directly been taken from their publication.

works (MTCNN) [189]. This is followed by horizontal flipping, and feature extraction by four separate networks. Feature-level fusion is performed by concatenation of features obtained for the original and flipped image, followed by concatenation of all features from different networks. $l_2$ normalization is performed on the concatenated feature vector, followed by classification using Cosine distance. The CNN architecture used in the proposed model is given in Figure 4-3(a).

**(ii) ByteFace:** Proposed by a team from Bytedance Inc., China, ByteFace uses an ensemble of three CNNs for performing disguised face recognition. For detection and alignment, the algorithm uses a mixture of co-ordinates provided with the DFW dataset and MTCNN. Three CNNs are trained with (i) modified center loss and Cosine similarity [173], (ii) joint Bayesian similarity, and (iii) sphere face loss [82] with joint Bayesian similarity, respectively. A linear weighted combination of scores obtained via the three models is used for performing the final classification. The CASIA WebFace [181] dataset is also used for training the proposed model.

**(iii) Deep Disguise Recognizer Network (DDRNET) [70]:** A team from West Virginia University, USA presented the DDRNET model. The name of the model was later changed to Deep Disguise Recognizer by the authors. Faces are cropped using the co-ordinates provided with the dataset, which is followed by pre-processing via whitening. An Inception network [156] along with Center loss [173] is trained on the pre-processed images, followed by classification using a similarity metric.

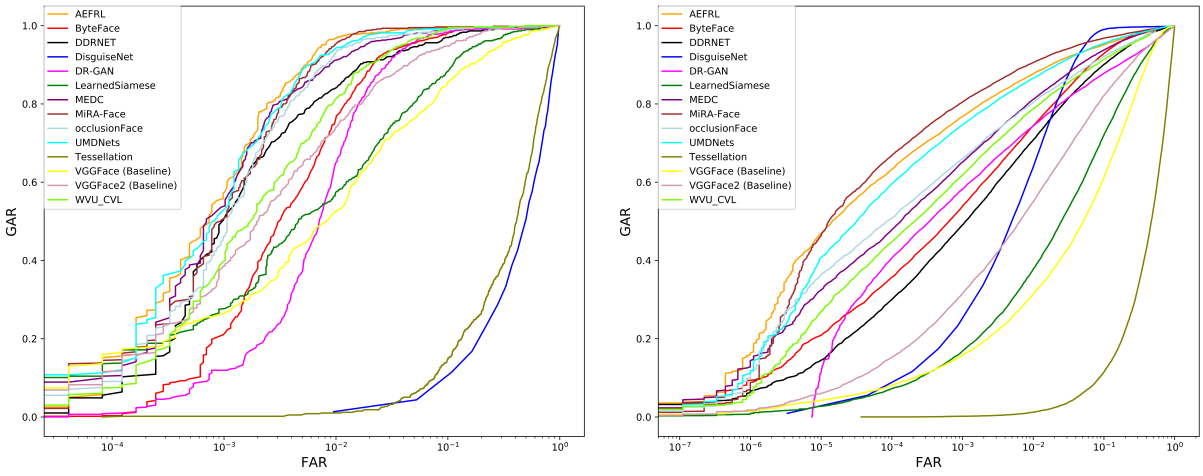
**(iv) DisguiseNet (DN) [115]:** Submitted by a team from the Indian Institute of Technology, Ropar, DisguiseNet performs face detection using the facial co-ordinates provided with the dataset. A Siamese network is built using the pre-trained VGG-Face [113], which is fine-tuned with the DFW dataset. Cosine distance is applied for performing classification of the learned features.

**(v) DR-GAN:** Proposed by a team from Michigan State University, USA, the framework performs face detection and alignment on the input images using MT-CNN [189]. This is followed by feature extraction using the Disentangled Representation learning-Generative Adversarial Network (DR-GAN) [160]. Classification is performed using Cosine distance.

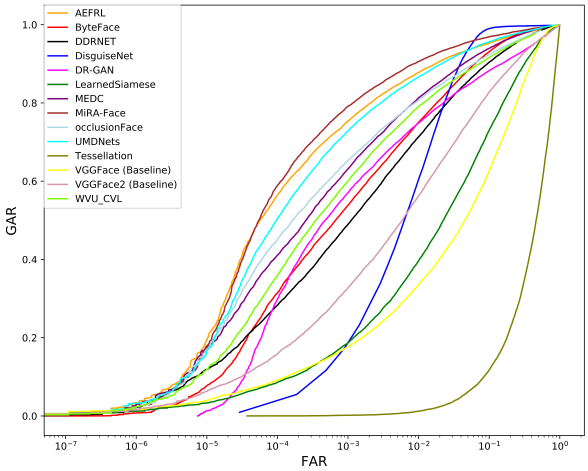
**(vi) LearnedSiamese (LS):** A team from the Computer Vision Center, Universitat AutÃšnoma de Barcelona, Spain proposed LearnedSiamese. Facial co-ordinates provided with the dataset are used for performing face detection, followed by learning a Siamese Neural Network for disguised face recognition.

**(vii) Model Ensemble with Different CNNs (MEDC):** MEDC is proposed by a team from the Northeastern University, USA. Face detection is performed using MTCNN followed by 2-D alignment. An ensemble of three CNNs is used for performing the given task of disguised face recognition. The algorithm utilizes a Center face model [173], Sphere face model [82], and a ResNet-18 model [47] trained on the MS-Celeb-1M dataset [46]. Since MS-Celeb-1M dataset also contains images taken from the Internet, mutual exclusion is ensured with the test set of the DFW dataset. Classification is performed using Cosine distance for each network, the average of which is used for computing the final result.

**(viii) MiRA-Face [187]:** Submitted by a team from the National Taiwan University, MiRA-Face uses a combination of two CNNs for performing disguised face recognition. It treats aligned and unaligned images separately, thereby using a context-switching technique for a given input




(a) Protocol-1 (b) Protocol-2




(c) Protocol-3

Figure 4-5: ROC curves of all participants along with the baseline results on protocol-1 (impersonation), protocol-2 (obfuscation), and protocol-3 (overall) of the DFW dataset.

image. Images are aligned using the co-ordinates provided with the dataset along with MTCNN and Recurrent Scale Approximation (RSA) [85]. Features learned by the CNNs are directly used for classification. Figure 4-4 presents a diagrammatic representation of the proposed model.

**(ix) OcclusionFace:** A team from ZJU, China proposed the OcclusionFace framework. MT-CNN [189] is used to perform face landmark detection and alignment based on five facial landmarks. ResNet-28 [47] is used for performing classification. The model is first pre-trained on the CASIA Webface dataset [181] followed by fine-tuning on the DFW dataset.

**(x) Tessellation:** Proposed by a team from Tessellate Imaging, India, Tessellation uses a Siamese network with triplet loss. Facial co-ordinates provided with the dataset are used for performing pre-processing, followed by training of the Siamese network. The final layer of the model learns a distance metric which returns a score between 0-1 for a given pair of images.

**(xi) UMDNets [7]:** Proposed by a team from University of Maryland, USA, its name was later modified to 'DCNN-based approach'. Face detection is performed by the All-in-One network [123], followed by alignment using the detected keypoints. Feature extraction is performed using two networks, followed by independent score computation. Classification is performed by averaging the scores obtained via the two feature sets. Figure 4-3(b) presents the training and testing pipeline of the proposed model.

**(xii) WVU_CL:** Submitted by a team from West Virginia University, USA, WVU_CL uses the face co-ordinates provided with the dataset along with MT-CNN [189] for face alignment. The aligned images are provided to a CNN architecture for performing classification using a softmax classifier.

### 4.4.3 Results

Tables 4.5-4.7 and Figure 4-5 present the Receiver Operating Characteristic (ROC) curves of the above mentioned models for all three protocols. Along with the submissions, the performance of VGG-Face [113] with Cosine distance is also tabulated as baseline. The performance of each model is reported in terms of Genuine Acceptance Rate (GAR) at 1% False Acceptance Rate (FAR) and 0.1% FAR. Results for each protocol are given in detail below:

**Results on Protocol-1 (Impersonation):** Figure 4-5(a) presents the ROC curves for all the submissions, and Table 4.5 presents the GAR corresponding to two FAR values. It can be observed that for the task of impersonation, AEFRL outperforms other algorithms at both the FARs by achieving 96.80% and 57.64% at 1% FAR and 0.1%FAR, respectively. A difference of around 40% is observed between the accuracies at both the FARs, which suggests that for scenarios having stricter authorized access, further improved performance is required. The second best performance is reported by MiRA-Face which presents a verification accuracy of 95.46% and 51.09%,

Table 4.5: Verification accuracy (%) of the participants and baseline performance on protocol-1 (impersonation).

| Algorithm | GAR | |
|---|---|---|
| | @1%FAR | @0.1%FAR |
| **AEFRL** | **96.80** | **57.64** |
| ByteFace | 75.53 | **55.11** |
| DDRNET | 84.20 | 51.26 |
| DenseNet + COST [3] | 92.10 | **62.20** |
| DisguiseNet | 1.34 [4] | 1.34 [5] |
| DR-GAN | 65.21 | 11.93 |
| LearnedSiamese | 57.64 | 27.73 |
| MEDC | 91.26 | **55.46** |
| **MiRA-Face** | **95.46** | 51.09 |
| OcclusionFace | 93.44 | 46.21 |
| Tessellation | 1.00 | 0.16 |
| **UMDNets** | **94.28** | 53.27 |
| VGGFace (Baseline) | 52.77 | 27.05 |
| VGGFace2 (Baseline) | 73.94 | 38.48 |
| WVU_CL | 81.34 | 40.00 |

respectively. At 0.1%FAR, MEDC performs second best and achieves an accuracy of 55.46%. All three algorithms utilize MT-CNNs for face detection and alignment before feature extraction and classification.

**Results on Protocol-2 (Obfuscation):** Figure 4-5(b) presents the ROC curves for the obfuscation protocol, and Table 4.6 summarizes the verification accuracies for all the models, along with the baseline results. MiRA-Face achieves the best accuracy of 90.65% and 80.56% for the two FARs. It outperforms other algorithms by a margin of at least 2.8% for GAR@1%FAR and 2.5% for GAR@0.1%FAR. As compared to the previous protocol (impersonation), the difference in the verification accuracy at the two FARs is relatively less. While further improvement is required, however, this suggests that recognition systems suffer less in case of obfuscation, as compared to impersonation at stricter FARs.

**Results on Protocol-3 (Overall):** Table 4.7 presents the GAR values of all the submissions, and Figure 4-5(c) presents the ROC curves for the third protocol. As with the previous protocol, MiRA-

[3]Not part of DFW competition
[4]GAR@0.95%FAR
[5]The smallest FAR value is 0.95%FAR for DisguiseNet.

Table 4.6: Verification accuracy (%) of the participants and baseline performance on protocol-2 (obfuscation).

| **Algorithm** | **GAR** | |
|---|---|---|
| | **@1%FAR** | **@0.1%FAR** |
| **AEFRL** | **87.82** | **77.06** |
| ByteFace | 76.97 | 21.51 |
| DenseNet + COST [3] | **87.10** | 72.10 |
| DDRNET | 71.04 | 49.28 |
| DisguiseNet | 66.32 | 28.99 |
| DR-GAN | 74.56 | 58.31 |
| LearnedSiamese | 37.81 | 16.95 |
| MEDC | 81.25 | 65.14 |
| **MiRA-Face** | **90.65** | **80.56** |
| OcclusionFace | 80.45 | 66.05 |
| Tessellation | 1.23 | 0.18 |
| **UMDNets** | **86.62** | **74.69** |
| VGGFace (Baseline) | 31.52 | 15.72 |
| VGGFace2 (Baseline) | 54.86 | 31.55 |
| WVU_CL | 78.77 | 61.82 |

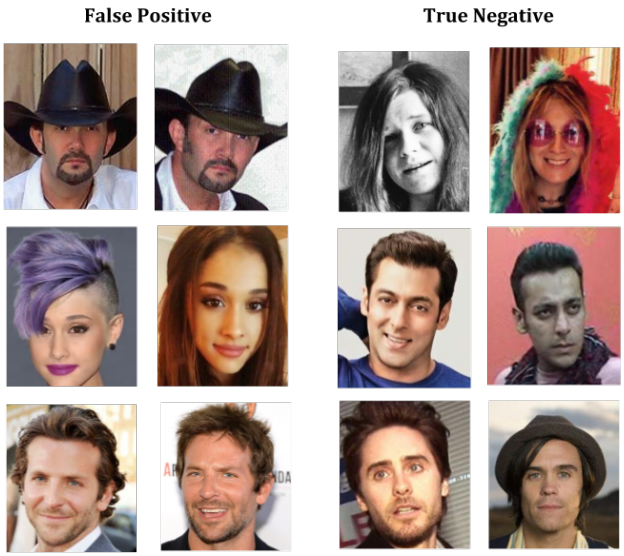



Figure 4-6: Sample False Positive and True Negative pairs reported by a majority of submissions for protocol-1 (impersonation). False Positive refers to the case where an algorithm incorrectly classifies a pair as genuine, and True Negative refers to the case where two samples of different identities are correctly classified as imposters.

Face outperforms other algorithms by a margin of at least around 3%. An accuracy of 90.62% and 79.26% is reported by the model for 1% and 0.1%FAR.

Table 4.7: Verification accuracy (%) of the participants and baseline performance on protocol-3 (overall).

| **Algorithm** | **GAR** | |
|---|---|---|
| | **@1%FAR** | **@0.1%FAR** |
| **AEFRL** | **87.90** | **75.54** |
| ByteFace | 75.53 | 54.16 |
| DenseNet + COST [3] | **87.60** | 71.50 |
| DDRNET | 71.43 | 49.08 |
| DisguiseNet | 60.89 | 23.25 |
| DR-GAN | 74.89 | 57.30 |
| LearnedSiamese | 39.73 | 18.79 |
| MEDC | 81.31 | 63.22 |
| **MiRA-Face** | **90.62** | **79.26** |
| OcclusionFace | 80.80 | 65.34 |
| Tessellation | 1.23 | 0.17 |
| **UMDNets** | **86.75** | **72.90** |
| VGGFace (Baseline) | 33.76 | 17.73 |
| VGGFace2 (Baseline) | 56.22 | 32.68 |
| WVU_CL | 79.04 | 60.13 |

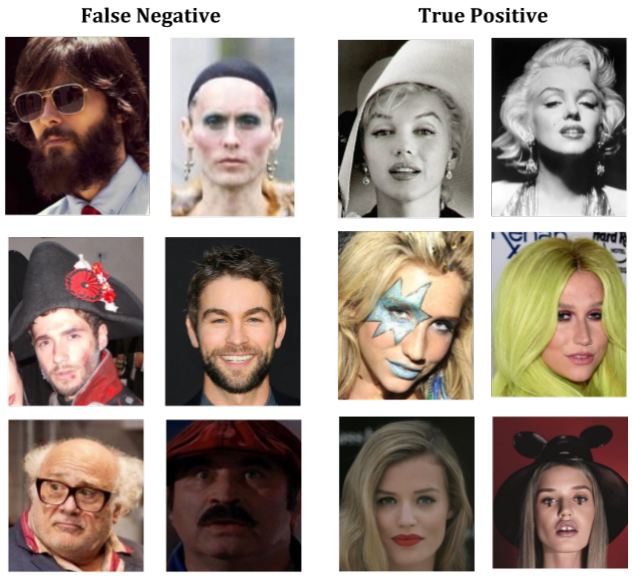



Figure 4-7: Sample False Negative and True Positive pairs reported by a majority of submissions for protocol-2 (obfuscation). False Negative refers to the case where a pair of images are incorrectly classified as an imposter pair, while True Positive refers to the scenario where a pair of images are correctly classified as a genuine pair.

Other than the DFW competition submissions, Suri *et al.* [154] proposed a novel COST (Color (CO), Shape (S), and Texture (T)) based framework for performing disguised face recognition. COST learns different dictionaries for Color, Shape, and Texture, which are used for feature extraction, along with the deep learning based model, DenseNet [56]. Final output is computed via classifier level fusion of the deep learning and dictionary learning models. The performance of the proposed DenseNet + COST algorithm has also been tabulated in Tables 4.12 - 4.14.

Figs. 4-6 - 4-7 demonstrate sample images of the DFW dataset correctly classified or misclassified by almost all the submissions. Figure 4-6 presents False Positive and True Negative samples for protocol-1 (impersonation). Upon analyzing the False Positive samples, it can be observed that all pairs have similar lower face structure, which might result in algorithms incorrectly classifying them as the same subject. Moreover, external disguises such as the cowboy hat (first pair) might also contribute to the misclassification. For protocol-2 (obfuscation), Figure 4-7 presents sample False Negative and True Positive pairs common across almost all submissions. It is interesting to observe that in the False Negative pairs, disguise results in modification of face structure and textural properties. Coupled with obfuscation of face and pose variations, the problem of disguised face recognition is rendered further challenging.

## 4.5 DFW2018 Dataset: Easy, Medium, and Hard Degree of Difficulty

In order to further analyze the DFW dataset, and study the problem of disguised faces in the wild, the DFW dataset has been partitioned into three sets: (i) easy, (ii) medium, and (iii) hard. The *easy* partition contains pairs of face images which are relatively easy to classify by a face recognition system, the *medium* set contains pairs of images which can be matched correctly by a majority of face recognition systems, while the *hard* partition contains image pairs with high matching difficulty. In literature, a similar partitioning was performed for the Good, Bad, and Ugly (GBU) face recognition challenge [116], where a subset of FRVT 2006 competition data [117] was divided into the three sets. The GBU challenge contained data captured over an academic year, in constrained settings with frontal face images having minimal pose or appearance variations.

Table 4.8: Number of *easy*, *medium*, and *hard* pairs for 1% and 0.1% FAR. TP and TN refer to True Positive (Genuine) and True Negative (Imposter), respectively.

| FAR | Number of | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Easy | | | Medium | | | Hard | | |
| | TP | TN | Total | TP | TN | Total | TP | TN | Total |
| 1% | 11,544 | 8,878,599 | 8,890,143 | 789 | 106,398 | 107,187 | 1,564 | 67,435 | 68,999 |
| 0.1% | 9,461 | 9,034,109 | 9,043,570 | 1,138 | 11,534 | 12,672 | 3,298 | 6,789 | 10,087 |

This section analyzes the DFW dataset containing data captured in unconstrained scenarios with variations across disguise, pose, illumination, age, and acquisition device.

The top-3 performing algorithms of the DFW competition have been used for partitioning the dataset, that is, AERFL, MiRA-Face, and UMDNets. The performance of the three algorithms is used for dividing the test set of the DFW dataset into three components: (i) easy, (ii) medium, and (iii) hard. *Easy* samples correspond to those pairs which were correctly classified by all three algorithms, and are thus easy to classify. *Medium* samples were correctly classified by any two algorithms, while the *hard* samples were correctly classified by only one algorithm, or mis-classified by all the algorithms, and thus are the most challenging component of the dataset. It is ensured that the partitions are disjoint, and samples belonging to one category do not appear in another category.

Table 4.16 presents the number of easy, medium, and hard pairs at different False Accept Rates of 1% and 0.1%. At 1%FAR, 11,544 genuine pairs are correctly classified as True Positive, while 8,878,599 imposter pairs are correctly classified as True Negative by all three techniques. This results in a total of 8,890,143 *easy* pairs, signifying that the total number of *easy* samples are highly dominated by the imposter pairs. In comparison, at 0.1%FAR, the total number of *easy* pairs increase to 9,043,570. It is interesting to observe that this increase is primarily due to the increased number of *easy* imposters at the lower FAR. Since at lower FARs, more pairs are classified as imposters, it leads to an increased number of *easy* pairs. Intuitively, at a stricter threshold of 0.1%FAR, one would expect the number of *easy* genuine samples to reduce. This trend is observed in Table 4.16, where the number of genuine pairs reduces from 11,544 at 1%FAR to 9,461 at 0.1%FAR.

The opposite trend is observed for the *hard* partition, where the total number of *hard* pairs reduces at 0.1%FAR, as compared to 1%FAR, however, the number of genuine samples increases.

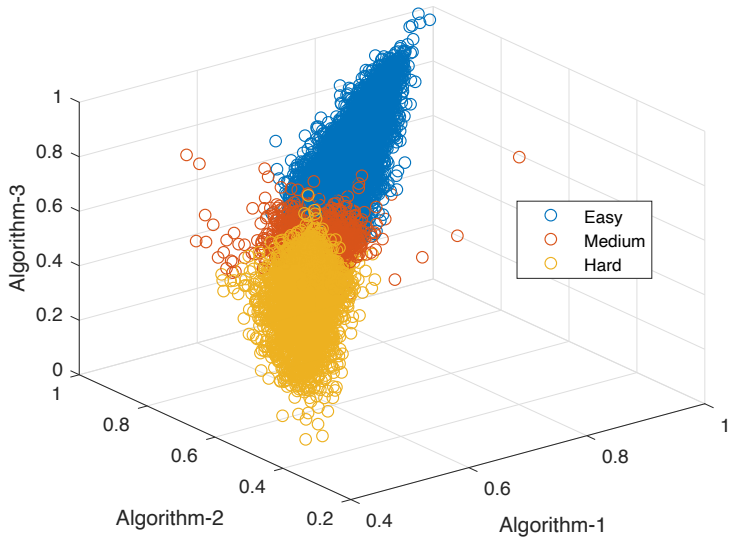



Figure 4-8: Score distribution of the genuine pairs at 0.01% FAR, in terms of three levels of difficulty: easy, medium, and hard.

The last three columns of Table 4.16 can be analyzed in order to observe this effect. At 1%FAR, the number of *hard* genuine pairs, that is, samples which are classified correctly by at most one algorithm is 1,564, while at 0.1%FAR it is 3,298. This implies that at a stricter FAR of 0.1%, more genuine samples were misclassified by all three algorithms. However, the number of *hard* imposter samples drops from 67,435 to 6,789 at a lower FAR. A similar trend is observed for the *medium* partition, wherein a total of 107,187 and 12,672 samples were correctly classified by any two algorithms at 1% and 0.1%FAR, respectively.

Figure 4-8 presents the score distribution of the genuine samples across the three categories of *easy*, *medium*, and *hard* at 0.1% FAR. The *easy* and *hard* samples occupy opposite ends of the distribution, while the *medium* category corresponds to a dense block between the two. Figure 4-9 presents sample *easy* and *hard* pairs of the DFW dataset at 0.1% FAR. The first row corresponds to *easy* genuine pairs, that is, genuine pairs correctly classified by all three top performing algorithms. Most of these pairs contain images with no pose variations ((i)-(ii)) or *similar* pose variations across images of the pairs ((iii)-(iv)). It can also be observed that most of these pairs are of the normal and validation images of the dataset, with minimal or no disguise variations. Images which involve disguise in terms of hair variations or hair accessories with minimal change in the face region are also viewed as *easy* pairs by the algorithms. Since in such cases, the face region remains

(a) *Easy* Genuine Samples

(b) *Hard* Genuine Samples

(c) *Hard* Imposter Samples

Figure 4-9: Sample *easy* and *hard* pairs of the DFW dataset.

unchanged, algorithms are often able to correctly classify such samples with ease. This observation is further substantiated by the *hard* genuine samples (Figure 4-9(b)). Most of the samples which were not correctly classified by any of the top algorithms contain occlusions in the face region. A large majority of genuine samples misclassified have occlusions near the eye region. All the pairs demonstrated in Figure 4-9(b) have at least one sample with occluded eye region. Effect of occlusion can also be observed in the *hard* imposter samples (Figure 4-9(c)), that is, imposters which were not correctly classified by either of the top-3 performing algorithms. Large variations due to heavy make-up, similar hair style or accessories, coupled with covariates of pose, occlusion, illumination, and acquisition device further make the problem challenging. It is our belief that in order to develop robust face recognition systems invariant to disguises, research must focus on addressing the *hard* pairs, while ensuring high performance on the *easy* pairs as well.

## 4.6 Disguised Faces in the Wild 2019 Dataset

The Disguised Faces in the Wild 2019 (DFW2019) dataset contains 3840 face images of 600 subjects. The images are collected from the Internet using relevant keywords from search engines,

Table 4.9: Statistics of the DFW2019 dataset.

| Image Variation | Number of | |
|---|---|---|
| | Subjects | Images |
| **Bridal** | 100 | 200 |
| **Plastic surgery** | 250 | 500 |
| **Other** | 250 | 3140 |
| **Total** | 600 | 3840 |

thereby demonstrating variability in terms of pose, illumination, resolution, acquisition mode, and disguise accessories. Other than images with external accessories such as hats, caps, beard, and sunglasses, the DFW2019 dataset also contains a subset of images having variations due to plastic surgery and bridal make-up. Broadly, the DFW2019 dataset contains two types of images: (i) subjects having before-after images with variations due to plastic surgery or bridal make-up, and (ii) subjects having unconstrained disguise variations due to occlusions or make-up, along with a *normal*, *validation*, and multiple *impersonator* images. Table 4.9 presents the statistics of the DFW2019 dataset. The dataset contains:

- 200 **bridal** images of 100 subjects, where each subject has two images corresponding to before and after applying bridal make-up,
- 500 **plastic surgery** images of 250 subjects, where each subject has two images corresponding to before and after the plastic surgery procedure,
- 3140 images of 250 subjects, where each subject contains a **validation** and a **normal** image, which corresponds to frontal and non-disguised high resolution images having good illumination. Each subject also contains a set of **disguised** images and a set of **impersonator** images (different subjects which appear to intentionally/unintentionally impersonate the subject). Figure 4-10 presents sample impersonator pairs.

The dataset will be made available for the research community[6]. Four protocols have also been defined for evaluations on the DFW2019 dataset. The following subsection elaborates upon each of the protocols.

[6]http://iab-rubric.org/resources.html

(a) (b)

Figure 4-10: Sample impersonator pairs created from the IIIT-Delhi Disguise dataset [26]. An individual can often use disguise accessories to impersonate another individual.

### 4.6.1 Protocols for Evaluation

Four verification protocols have been presented for evaluating face recognition algorithms on the DFW2019 dataset. Continuing from the DFW2018 competition [143], two protocols are: (i) Impersonation and (ii) Obfuscation, while the remaining two correspond to (iii) Plastic Surgery and (iv) Overall. The following paragraphs present each protocol in detail, along with the description of genuine and imposter sets.

**Protocol 1 - Impersonation:** This protocol aims to assess a face recognition system under the effect of impersonation. Here, the genuine set consists of the normal-validation image pair of the same subject, and the imposter set consists of normal-impersonator pair, disguise-impersonator pair, and validation-impersonator pair of the same subject. In this protocol, there exist 250 genuine and 7,431 imposter pairs.

**Protocol 2 - Obfuscation:** This protocol focuses on evaluating a face recognition system under intentional or unintentional disguise variations of genuine users. Here, the genuine set corresponds to the normal-disguise, validation-disguise, and $\text{disguise}_1$-$\text{disguise}_2$ image pairs of the same subject, along with the before-after bridal make-up images. The imposter set contains cross-subject pairs, where the disguised, normal, and validation images of one subject are paired with the disguised, normal, and validation images of another subject. Moreover, cross-subject before-after pairs for the bridal make-up set also constitute the imposter set. In total, this protocol contains 10,267 genuine and 2,802,011 imposter pairs.

**Protocol 3 - Plastic Surgery:** This protocol is specifically targeted towards evaluating a face recognition system against changes in facial features due to plastic surgery. Here, the before-after images of subjects who have undergone plastic surgery are utilized. The genuine set (250

Table 4.10: Baseline results on the DFW2019 dataset. GAR is reported for the specified FAR values.

| Protocol | Model | 0.1% FAR | 0.01%FAR |
|---|---|---|---|
| **P-1** | ResNet-50 | 47.6 | 38.4 |
| | LightCNN-29v2 | 74.4 | 51.2 |
| **P-2** | ResNet-50 | 35.3 | 16.4 |
| | LightCNN-29v2 | 55.5 | 36.9 |
| **P-3** | ResNet-50 | 46.4 | 22.4 |
| | LightCNN-29v2 | 69.2 | 47.2 |
| **P-4** | ResNet-50 | 35.9 | 16.8 |
| | LightCNN-29v2 | 55.7 | 36.5 |

pairs) contains the before-after images of the same subject, while the imposter set (124,500 pairs) contains cross-subject before-after images.

**Protocol 4 - Overall:** The overall protocol attempts to evaluate a face recognition system on the entire DFW2019 dataset. Here, the genuine set contains a combination of all the images in the genuine sets of Protocols 1-3. That is, the genuine set contains the normal-validation (Protocol-1), validation-disguise, normal-disguise, $\text{disguise}_1$-$\text{disguise}_2$, before-after bridal make-up (Protocol-2), and before-after plastic surgery (Protocol-3) image pairs. The imposter set also contains a combination of the imposter pairs across Protocols 1-3. That is, the imposter set contains normal-impersonator, disguise-impersonator, validation-impersonator (Protocol-1), cross-subject imposters, cross-subject before-after bridal make-up (Protocol-2), and cross-subject before-after plastic surgery (Protocol-3) pairs.

## 4.7 Baseline Results

For all the protocols, baseline results have been computed using two pre-trained state-of-the-art deep learning based face recognition models. ResNet-50[7] [47] (pre-trained on the large-scale VGG-Face2 [15] and MS-Celeb-1M [46] datasets) and LightCNN-29v2[8] [177] (pre-trained on the large-scale CASIA-WebFace [181] and MS-Celeb-1M [46] datasets) have been used for evaluation. Pre-trained models were used as is, without any additional training. Detected and cropped face images were provided to the network, followed by feature extraction, and Cosine similarity

[7]https://github.com/cydonia999/VGGFace2-pytorch
[8]https://github.com/AlfredXiangWu/LightCNN

based classification. Face detection was performed using the Tiny Face detector [55], followed by manual detection of the false negative faces. The extracted embeddings were of dimension 2048 and 256 for ResNet-50 and LightCNN-29v2, respectively. Genuine Acceptance Rate (GAR) is reported for fixed False Acceptance Rates (FARs), which form the baselines for the DFW2019 dataset.

Table 4.10 presents the baseline results obtained for the DFW2019 dataset using the two networks: Resnet-50 and LightCNN-29v2. Results have been tabulated for two FARs: 0.1% and 0.01% for all the protocols (protocol 1-4). LightCNN-29v2 consistently outperforms the ResNet-50 model by achieving improved verification performance across all protocols and FARs.

## 4.8 Disguised Faces in the Wild 2019 Competition

The DFW2019 competition[9] was held in conjunction with the *International Workshop on Disguised Faces in the Wild* at the International Conference on Computer Vision (ICCV), 2019. Participants had to develop a face recognition model which is evaluated on the DFW2019 dataset.

Anonymized DFW2019 dataset was provided to the participants as the test set, and evaluation is performed on all four protocols. The training and testing partitions of the DFW2018 dataset [143] were also provided as the training and validation partition, respectively, for the competition. The DFW2019 dataset will be made publicly available for the research community. We believe that the DFW2019 dataset can help in enhancing the recognition performance for disguised faces, thereby improving the robustness of face recognition algorithms.

### 4.8.1 DFW2019 Competition: Submissions

The DFW2019 competition received over 100 registrations and 11 submissions from all over the world. Table 4.11 summarizes the affiliation of the different submissions received as part of this competition. Each submission is described in detail as follows:

**(i) ArcFace:** A team from the Imperial College London proposed using ArcFace [20] (Additive Angular Margin Loss) for recognizing disguised faces in the wild. The model incorporates a

[9] http://iab-rubric.org/DFW/2019Competition.html

Table 4.11: List of teams who participated in the DFW2019 competition.

| Algorithm | Team | Institution |
|---|---|---|
| A-1 | ArcFace | Imperial College London |
| A-2 | ArcFaceInter | Imperial College London |
| A-3 | ArcFaceIntra | Imperial College London |
| A-4 | ArcFaceIntraInter | Imperial College London |
| A-5 | FakeFace | ITMO University |
| A-6 | FakeFacev2 | ITMO University |
| A-7 | FEBNet | Indian Institute of Technology, Madras |
| A-8 | LightCNNDFW | Anonymous |
| A-9 | Mozart | Tech5.ai |
| A-10 | SEBNet | Indian Institute of Technology, Madras |
| A-11 | XuXu | Tech5.ai |

margin in the popularly used Softmax loss for deep learning based Convolutional Neural Networks. Facial co-ordinates are computed using the RetinaFace model [21].

**(ii) ArcFaceInter:** Submitted by a team from the Imperial College London, ArcFaceInter incorporates an additional term for enhancing the inter-class distance in the ArcFace [20] model. RetinaFace [21] is used for computing the facial co-ordinates and geometric alignment of images.

**(iii) ArcFaceIntra:** ArcFaceIntra incorporates an intra-class penalty to enhance class compactness into the ArcFace model. Submitted by a team from the Imperial College London, features are extracted from the ArcFaceIntra model for faces detected and aligned via the RetinaFace model [21].

**(iv) ArcFaceIntraInter:** ArcFaceIntraInter models both inter-class and intra-class variations during feature learning. Submitted by a team from the Imperial College London, ArcFaceIntraInter incorporates two additional terms in the ArcFace model [20] for increasing the inter-class distance and reducing the intra-class variations. Face detection and alignment is performed using the RetinaFace [21] model.

**(v) FakeFace:** Submitted by a team from the ITMO University, Russia, faces are detected and aligned with RetinaFace [21] and cropped to $112 \times 112$. A deep learning network is trained using the MS-Celeb-1M dataset [46] and the ArcFace loss [20]. The model is fine-tuned with Doppleganger Mining [147], Auxillary Embeddings [148], Embeddings Interpolations, and Priority Lists. An ensemble of three such networks is used for feature extraction.

**(vi) FakeFacev2:** Submitted by a team from the ITMO University, Russia, FakeFacev2 uses a combination of RetinaFace [21] and ArcFace [20] as backbone for recognizing disguised faces in the wild. Fine-tuning is performed on the MS-Celeb-1M dataset [46] with Doppleganger Mining [147], Auxillary Embeddings [148], Embeddings Interpolations, and Priority Lists. Evaluation is performed using an ensemble of three such networks.

**(vii) FEBNet:** A team from the Indian Institute of Technology, Madras proposed the FEBNet model. Detected faces provided with the dataset are used with an ensemble of SE-ResNet-50 (pre-trained on the MS-Celeb-1M dataset [46]) and Inception-ResNet-v1 (pre-trained on the VGGFace2 dataset [15]). Fine-tuning is performed using a combination of identity loss, triplet loss, and category loss. Decision is taken via score-level fusion and a re-ranking approach.

**(viii) LightCNNDFW:** A pre-trained LightCNN-29v2 [177] network has been fine-tuned in a Siamese manner. Binary cross-entropy loss is applied on the extracted features. Detected faces provided with the dataset are used, along with the *five-crop* data augmentation technique.

**(ix) Mozart:** Submitted by a team from Tech5.ai, Mozart uses the detected faces provided with the DFW2019 dataset. An ensemble of different ResNet models is used for feature extraction, followed by matching via the $l_2$-distance.

**(x) SEBNet:** SEBNet has been submitted by a team from the Indian Institute of Technology, Madras and utilizes an ensemble of deep learning networks. Two networks: InceptionNet-v3 (pre-trained on the MS-Celeb-1M dataset [46]) and SE-ResNet-50 (pre-trained on the VGGFace2 dataset [15]) are fine-tuned on the DFW2018 dataset. The trained networks are used for feature extraction, followed by Euclidean distance based score computation, score-level fusion, and a re-ranking algorithm.

**(xi) XuXu:** Submitted by a team from Tech5.ai, XuXu utilizes an ensemble of different ResNet models. The pipeline includes geometric alignment on the detected faces provided with the dataset, followed by feature extraction from the ensemble. Matching is performed using $l_2$-distance.

### 4.8.2 Results

For all the protocols, results are reported in the form of Genuine Acceptance Rate (GAR) for the specified False Acceptance Rates (FAR). Baseline results have been reported using the LightCNN-

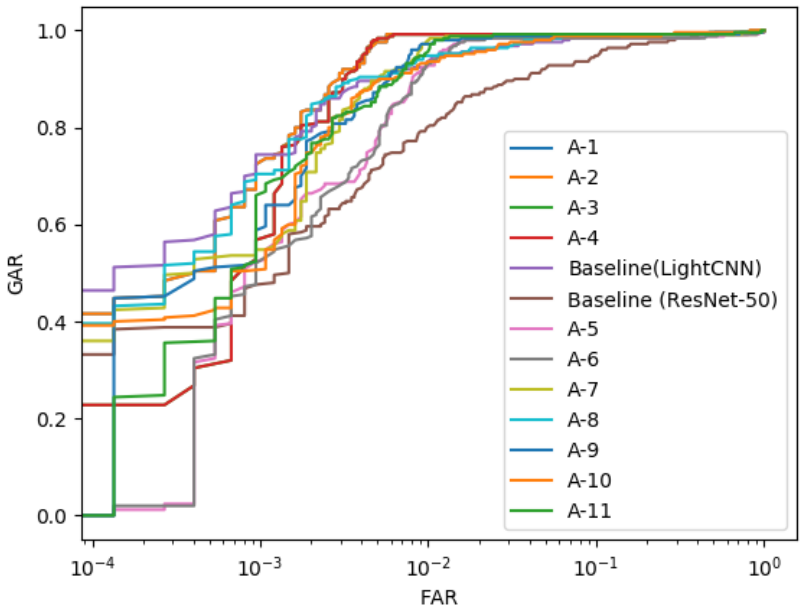
Table 4.12: Verification accuracy (%) on the proposed DFW2019 dataset for the Impersonation protocol (Protocol-1). The table presents the performance of participants and the baseline results.

| **Algorithm** | **GAR** | |
|---|---|---|
| | **@0.1%FAR** | **@0.01%FAR** |
| A-1 | 72.4 | 44.8 |
| A-2 | 72.4 | 44.8 |
| A-3 | 56.8 | 17.6 |
| A-4 | 56.8 | 17.6 |
| A-5 | 52.4 | 1.2 |
| A-6 | 52.0 | 2.0 |
| A-7 | 54.8 | 42.4 |
| A-8 | 70.4 | 43.2 |
| A-9 | 58.8 | 44.8 |
| A-10 | 54.8 | 40.0 |
| A-11 | 66.0 | 24.4 |
| **Baseline (LightCNN)** | **74.4** | **51.2** |
| Baseline (ResNet-50) | 47.6 | 38.4 |

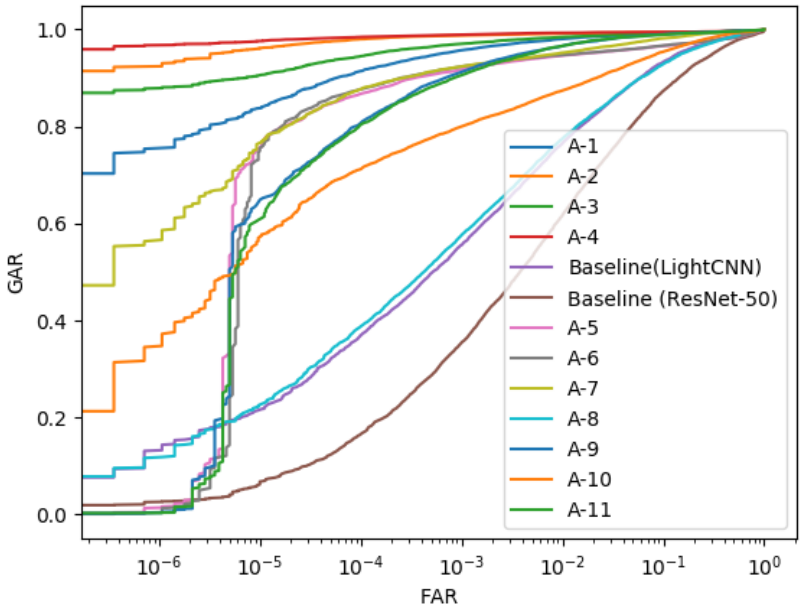
29v2 model [177] and the ResNet-50 model [47], with Cosine similarity based classification (Section 4.7). The following paragraphs elaborate upon the results obtained by for each protocol, including the submissions and the baseline results:

**(i) Protocol-1 (Impersonation):** Figure 4-11(a) contains the ROC curves for the baseline results and the submissions. Table 4.12 presents the GAR at 0.1% and 0.01% FAR for all the submissions. At both the FARs, the baseline performance of LightCNN-29v2 performs the best by achieving 74.4% and 51.2%, respectively. At both the FARs, A-1 (ArcFace) and A-2 (ArcFaceInter) perform second best with GARs of 72.4% and 44.8%, respectively. A drop of around 24% is observed between the verification performance at 0.1% and 0.01% FAR of LightCNN-29v2, suggesting the need for face recognition models to focus more on preventing impersonation based attacks.

**(ii) Protocol-2 (Obfuscation):** Figure 4-11(b) demonstrates the ROC curves for Protocol-2 (obfuscation), and Table 4.13 presents the GAR values at two specified FARs: 0.1% and 0.01%. A-4 (ArcFaceIntraInter) outperforms other techniques by reporting a GAR of 98.9% and 98.4% at 0.1% and 0.01% FAR, respectively. The second and third best performance are also obtained by variants of the ArcFace model. In Protocol-2 the variations observed between the GAR at 0.1% and 0.01% is less than that obtained in Protocol-2. The improved GARs at lower FARs further suggest that deep learning based face recognition models are able to handle variations due to obfuscation better,




(a) Protocol-1 (Impersonation)




(b) Protocol-2 (Obfuscation)

Figure 4-11: ROC curves on the DFW2019 dataset for Protocol-1 and Protocol-2. Images have been taken from the published manuscript [136].

that is, scenarios where a genuine user attempts to obfuscate their identity by means of an external accessory.

**(iii) Protocol-3 (Plastic Surgery):** Figure 4-12(a) presents the ROC curves for Protocol-3, that

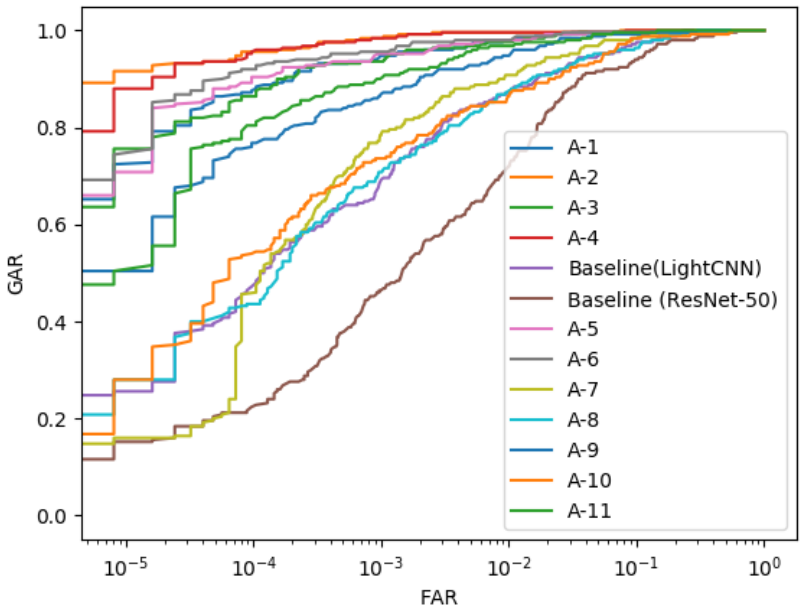
Table 4.13: Verification accuracy (%) on the proposed DFW2019 dataset for the Obfuscation protocol (Protocol-2). The table summarizes the performance of participants and the baseline results.

| **Algorithm** | **GAR** | |
|---|---|---|
| | **@0.1%FAR** | **@0.01%FAR** |
| A-1 | 95.7 | 91.4 |
| A-2 | 98.7 | 97.9 |
| A-3 | 97.0 | 94.4 |
| **A-4** | **98.9** | **98.4** |
| A-5 | 91.6 | 86.6 |
| A-6 | 92.3 | 87.7 |
| A-7 | 92.3 | 87.6 |
| A-8 | 57.5 | 38.6 |
| A-9 | 91.1 | 80.5 |
| A-10 | 80.0 | 71.2 |
| A-11 | 90.5 | 80.5 |
| Baseline (LightCNN) | 55.5 | 36.9 |
| Baseline (ResNet-50) | 35.3 | 16.4 |

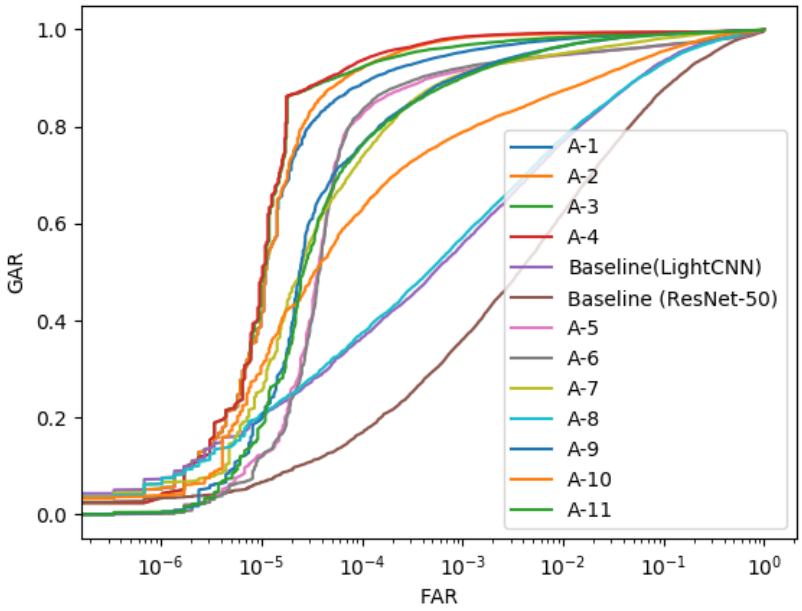
Table 4.14: Verification accuracy (%) for the Plastic Surgery protocol (Protocol-3). Results of the submissions and baseline performance computed using ResNet-50 and LightCNN-29v2 have been presented in the table.

| **Algorithm** | **GAR** | |
|---|---|---|
| | **@0.1%FAR** | **@0.01%FAR** |
| A-1 | 94.8 | 87.6 |
| **A-2** | **98.4** | **95.6** |
| A-3 | 93.6 | 86.4 |
| **A-4** | **98.4** | **95.6** |
| A-5 | 95.2 | 90.4 |
| A-6 | 95.6 | 92.0 |
| A-7 | 78.8 | 47.6 |
| A-8 | 70.8 | 43.6 |
| A-9 | 86.8 | 76.8 |
| A-10 | 73.6 | 54.0 |
| A-11 | 90.0 | 81.2 |
| Baseline (LightCNN) | 69.2 | 47.2 |
| Baseline (ResNet-50) | 46.4 | 22.4 |

is, variations brought in the face due to the plastic surgery procedure. Table 4.14 also presents the GAR values obtained at the specified FARs of 0.1% and 0.01% for all the submissions and baseline results. Best performance of 98.4% and 95.6% is obtained via A-2 (ArcFaceInter) and




(a) Protocol-3 (Plastic Surgery)




(b) Protocol-4 (Overall)

Figure 4-12: ROC curves on the DFW2019 dataset for Protocol-3 and Protocol-4.

A-4 (ArcFaceIntraInter) for 0.1% and 0.01% FAR, respectively. The second and third best performance are obtained by A-6 (FakeFacev2) and A-5 (FakeFace) submissions, wherein a difference of around 3% is observed at 0.1%FAR. High verification performance on both FARs demonstrate the effectiveness of the submissions for handling face recognition under variations due to plastic

Table 4.15: Verification accuracy (%) for the Overall protocol (Protocol-4). The table presents the performance of the participants and baseline results computed using ResNet-50 and LightCNN-29v2.

| Algorithm | GAR | |
|---|---|---|
| | **@0.1%FAR** | **@0.01%FAR** |
| A-1 | 95.2 | 88.6 |
| A-2 | 98.3 | 92.0 |
| A-3 | 96.7 | 92.1 |
| **A-4** | **98.4** | **93.6** |
| A-5 | 91.4 | 82.2 |
| A-6 | 92.1 | 83.1 |
| A-7 | 90.7 | 73.6 |
| A-8 | 57.1 | 37.4 |
| A-9 | 90.7 | 76.1 |
| A-10 | 78.8 | 62.8 |
| A-11 | 90.0 | 76.0 |
| Baseline (LightCNN) | 55.7 | 36.5 |
| Baseline (ResNet-50) | 35.9 | 16.8 |

surgery.

**(iv) Protocol-4 (Overall):** Protocol-4 evaluates the performance of a face recognition system on the entire DFW2019 dataset. Figure 4-12(b) presents the ROC curves of the submissions and baseline results, and Table 4.15 presents the GAR values obtained at 0.1% and 0.01% FAR, respectively. A-4 (ArcFaceIntraInter) achieves the highest performance on both the FARs: 98.4% and 93.6% at 0.1% and 0.01% FAR, respectively. This is followed by A-2 (ArcFaceInter) and A-3 (ArcFaceIntra) on both the FARs.

Overall, the DFW2019 competition received 11 submissions, all of which utilized deep learning based pre-trained networks. It is our belief that the availability of networks pre-trained on large datasets facilitates discriminative feature extraction, resulting in high performance.

## 4.9 DFW2019 Dataset: Easy, Medium, and Hard Pairs

Based on the degree of difficulty of verifying a pair of face images, the DFW2019 dataset is divided into three components: *easy*, *medium*, and *hard*. This section presents an analysis of the dataset along the above mentioned components. The *easy* partition contains those image pairs which are

Table 4.16: Total *easy*, *medium*, and *hard* pairs at 0.01% FAR. *Easy* refers to the number of pairs correctly classified as TP (True Positive)/TN (True Negative). *Medium* refers to the number of pairs correctly classified as TP/TN by two algorithms, while *Hard* refers to the number of TP/TN pairs correctly classified by at most one algorithm.

| | **Genuine (TP)** | **Imposter (TN)** | **Total** |
|---|---|---|---|
| **Easy** | 7,743 | 2,933,312 | 2,941,055 |
| **Medium** | 1,595 | 445 | 2,040 |
| **Hard** | 1,429 | 185 | 1,614 |

relatively easy to correctly verify by face recognition algorithms. On the other hand, the *hard* partition contains those image pairs which are harder to verify by face recognition algorithms. Division of the DFW2019 dataset in easy, medium, and hard categories is similar in concept to the partitioning of the DFW2018 dataset [143], as well as the Good, Bad, and Ugly components of the FRVT 2006 competition dataset [117].

For the DFW2019 dataset, the results obtained by the top-3 teams for the Overall protocol (protocol-4) have been utilized to create the (i) *easy*, (ii) *medium*, and (iii) *hard* partition. As observed from Table 4.15, the top-3 teams correspond to: (i) A-4 (ArcFaceIntraInter), (ii) A-6 (FakeFacev2), and (iii) A-7 (Mozart). For the DFW2019 dataset, the *easy* partition corresponds to the image pairs correctly classified by all three algorithms. The *medium* partition contains pairs of face images which have been correctly classified by any two of the top-3 submitting teams, while the *hard* partition contains image pairs which have been classified correctly by any one algorithm, or have been incorrectly matched by all three algorithms. The partitioning of the DFW2019 dataset has been performed for both genuine and imposter pairs, and mutual exclusion has been ensured across the three partitions.

Table 4.16 presents the count of the easy, medium, and hard pairs for the DFW2019 dataset at 0.01%FAR. For the genuine set, 7,743 pairs belong to the *easy* category which were correctly matched by the top three teams. On the other hand, 1,429 pairs correspond to the genuine *hard* partition which were incorrectly classified by at least two of the top three teams (almost 14% of the entire genuine set). Figure 4-13 presents a Venn Diagram demonstrating the number of mis-classifications of genuine pairs from the DFW2019 dataset. It can be observed that 603 pairs were mis-classified by all top-3 teams, which form a part of the hard partition for the DFW2019 dataset. In total, the medium and hard partitions correspond to 3,654 pairs of face images from the

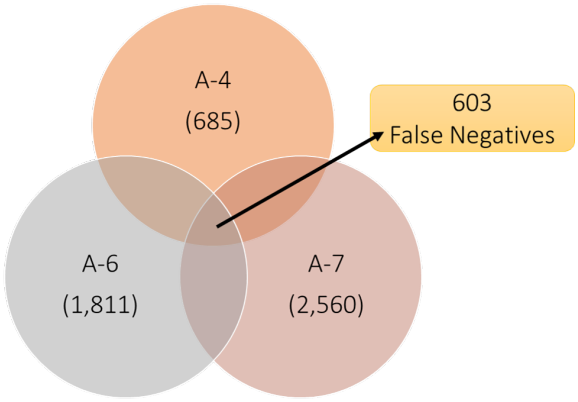



Figure 4-13: Venn diagram demonstrating the number of mis-classifications of the genuine pairs by the top-3 teams (A-4: ArcFaceIntraInter, A-6: FakeFacev2, A-7:Mozart) at 0.01% FAR. The common region (603 samples) is a subset of the *hard* samples which were mis-classified by all algorithms.

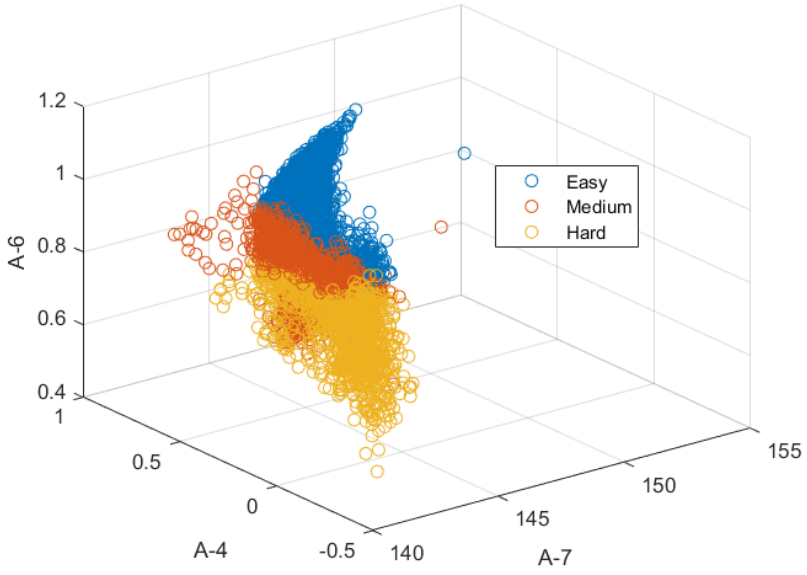



Figure 4-14: Scatter plot of the scores obtained by the top-3 teams for the Easy, Medium, and Hard pairs of the DFW2019 dataset.

DFW2019 dataset. Figure 4-14 presents the scores obtained by the top-3 algorithms for the three partitions. Scores for the easy and hard sets of the DFW2019 dataset occupy opposite ends of the distribution, while scores corresponding to the medium partition are present in the middle.

In several law enforcement applications, face recognition systems are often required to operate under the strict threshold of 0% FAR. That is, no imposter pair should be incorrectly classified

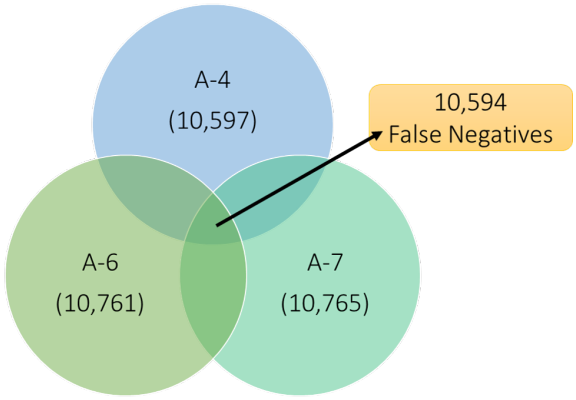



Figure 4-15: Venn diagram demonstrating the number of mis-classifications of the genuine pairs (True Positive samples) by the top-3 teams (A-4: ArcFaceIntraInter, A-6: FakeFacev2, A-7:Mozart) for 0 False Positives. The common region (10,594 samples) corresponds to a subset of samples which were mis-classified by all algorithms.

as a genuine pair (0 FAR) while correctly classifying the genuine set of images (high GAR). On the DFW2019 dataset, Figure 4-12(b) can be analyzed to observe very low performance at lower FARs for the overall protocol. Figure 4-15 presents a Venn Diagram for the number of incorrect classifications of the genuine set by the top-3 algorithms at 0% FAR. 10,594 pairs of images are mis-classified by all three algorithms, which corresponds to 98.39% of the total genuine samples. The reduced performance at lower FARs suggests the need for robust face recognition systems applicable to critical law enforcement applications. It is our belief that moving forward, face recognition algorithms should focus on further reducing the number of hard pairs, while achieving high accuracy on the easy partition.

## 4.10 Summary

This research presents a novel Disguised Faces in the Wild dataset (referred to as the DFW or the DFW2018 dataset) and the Disguised Faces in the Wild 2019 (referred to as the DFW2019) dataset. The DFW2018 dataset contains 11,157 images pertaining to 1,000 identities with variations across different disguise accessories, while the DFW2019 dataset contains 3,840 images of 600 subjects. All images are collected from the Internet via relevant keyword searches on different search engines, thereby demonstrating wide variations with respect to pose, illumination, lighting, resolution, capturing device, and disguise accessories. A given subject may contain four types of

images: normal, validation, disguised, and impersonator. Out of these, normal and validation images are non-disguised frontal face images. Disguised images of a subject contain genuine images of the same subject with different disguises. Impersonator images correspond to images of different people who try to impersonate (intentionally or unintentionally) another subject. Additionally, the DFW2019 dataset contains variations due to different disguise accessories, and before-after images for plastic surgery and bridal make-up. To the best of our knowledge, these are the first disguised face datasets to provide impersonator images for different subjects. This research also presents pre-defined evaluation protocols and baseline results of two state-of-the-art deep learning based networks: LightCNN-29v2 [177] and ResNet-50 [47] for both datasets. The two datasets were released as part of the International Workshop on Disguised Faces in the Wild, held in conjunction with the Computer Vision and Pattern Recognition (CVPR) conference, 2018, and the International Conference on Computer Vision (ICCV), 2019, respectively. This research also summarizes the performance of the submissions received as part of the competitions, and analysis has also been performed by partitioning the datasets into three components: (i) easy, (ii) medium, and (iii) hard. Dedicated research in the direction of disguised face recognition could help in the development of robust face recognition systems, imperative for several real world applications. It is our hope that the DFW dataset would help facilitate research in this important yet less explored domain of face recognition. This research has been published in the IEEE Transactions on Biometrics, Behavior, and Identity Science; and the IEEE/CVF International Conference on Computer Vision Workshops, 2019. All the images have been taken from the above published manuscripts [136, 143].

# Chapter 5

# Disguised Resilient Face Verification

## 5.1 Introduction

Recently, the Los Angeles Police Department released the top phrases used to describe suspects[1]. Most of the keywords correspond to some form of accessory used to obscure one's face, such as *cap/hat, hoodie, mask,* or *wig*. Such accessories are often also used in day-to-day life, resulting in unintentional obfuscation of the facial region. Hidden biometric information (e.g. facial region) often causes challenges to a face recognition system [150], presenting the need to develop systems invariant to disguise variations.

Disguised face recognition refers to the task of matching face images with variations due to disguise accessories. It has wide-spread applicability in scenarios related to law-enforcement and surveillance. An automated recognition system often utilizes data captured via surveillance cameras such as CCTV cameras, resulting in low resolution facial regions (often less than $32 \times 32$). To the best of our knowledge, no research has focused on low resolution face recognition under disguise variations; an important yet unexplored problem. To this effect, this research proposes a novel Disguise Resilient framework for disguised face verification, applicable to low resolution facial images as well.

Broadly, the use of disguise accessories can either be to (i) obfuscate one's identity or (ii) impersonate another individual. Unintentional use of accessories such as sunglasses or scarves often obfuscate the face region [35, 45], rendering automated face recognition challenging. On the

[1]https://lasentinel.net/suspect-wore-a-hoodie.html

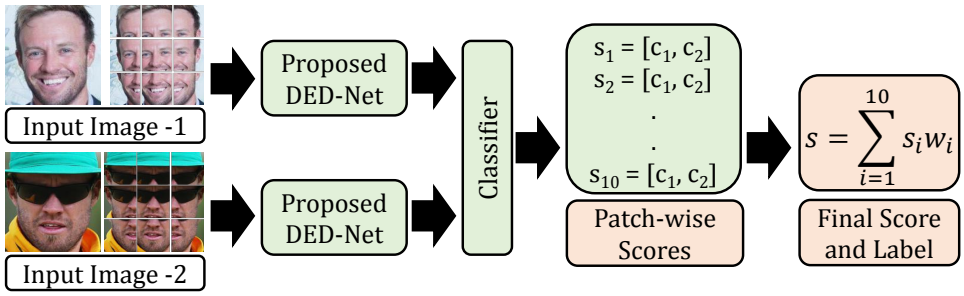



Figure 5-1: This research presents a novel Disguise Resilient (D-Res) framework for learning disguise invariant features. The framework utilizes the proposed DED-Net architecture for processing tessellated face images. Ten score pairs ($s_i = [c_i, c_2]$) are obtained via the classifier (one for each patch) for genuine and imposter classes. The scores are combined in a weighted manner (using weights $w_i$) to generate the final score $s$ for the given input image pair.

other hand, such accessories can often also be used to impersonate another individual, or appear as their look-alike. In such scenarios, a robust face recognition system should be able to correctly identify the individual as an imposter. Recent research in disguised face recognition has resulted in algorithms achieving promising results for the scenario of obfuscation. For example, in the recent Disguised Faces in the Wild 2019 competition [136], for the scenario of obfuscation, the top performing team achieved 98.9% and 98.4% Genuine Acceptance Rate (GAR) at 0.1% and 0.01% False Acceptance Rate (FAR), respectively [22]. On the other hand, at the same FAR values, the best performance of 74.4% and 51.2% GAR was obtained for the impersonation protocol [136]. The considerable low performance for the impersonation protocol suggests the challenging nature of the problem, and the need for dedicated research focus.

This research presents two-fold contributions: (i) a novel supervised Encoder-Decoder formulation, termed as the *DED-Net*, capable of learning disguise-invariant features from face images, and (ii) the proposed *Disguise Resilient (D-Res) framework* (Figure 5-1) for disguised face verification, applicable to low resolution face images as well. Global and local features are learned from the full face and different facial patches in order to learn an effective verification model. Extensive evaluation is performed on the two most recent and challenging datasets: the DFW2018 [143] and DFW2019 [136] datasets. This is also the first work presenting benchmark results and baselines for low resolution disguise face verification for three different resolutions: $32 \times 32$, $24 \times 24$, and $16 \times 16$. Experimental evaluation demonstrates the accuracy improvement achieved by the D-Res framework on multiple protocols for both the datasets.

## 5.2 Related Work

Automated disguised face recognition has garnered the attention of the research community over the past several years. Initial research focused on identifying the non-disguised facial patches, followed by the extraction of hand-crafted features invariant to disguise variations. For example, Ramanathan *et al.* [121] and Martinez *et al.* [92] utilized a combination of eigenspaces and Mahalanobis distance for disguised face recognition. Singh *et al.* [145] proposed using 2D log-polar Gabor features for disguised face verification, while Wright *et al.* [176] focused on extracting sparse features for faces with occlusions, to eliminate the effect of disguise on face recognition. Dhamecha *et al.* [26] proposed identifying non-disguised facial patches, and matching using a Local Binary Pattern based algorithm.

Recently, with the advent of representation learning algorithms, researchers have also tried to learn disguise-invariant representations for efficient face recognition. Moreover, the availability of real-world datasets and challenging competitions, such as the Disguised Faces in the Wild (DFW) competition series [136, 143], have further pushed the state-of-the-art for disguised face recognition. Smirnov *et al.* [148] proposed learning a face verification network using hard example mining and auxiliary embeddings. Zhang *et al.* [187], Subramaniam *et al.* [151], and Bansal *et al.* [7] proposed using ensembles of deep learning based face recognition systems for addressing disguised face verification. In 2019, Deng and Zafeiriou [22] proposed using the ArcFace loss, while modeling the inter-class and intra-class variations. Suri *et al.* [153] proposed *A-LINK*, an active learning based inter-domain knowledge algorithm for modeling the disguise variations in deep learning networks.

Despite the recent advances and improvement in disguised face verification, the state-of-the-art performance is substantially lower as compared to traditional face recognition. For example, the best reported results for the *Impersonation* protocol of the DFW 2019 dataset are around 78% [136], whereas over 99% verification performance has already been achieved on the widely-used Labeled Faces in the Wild (LFW) dataset [57]. Further, to the best of our knowledge, no research has focused on understanding the resolution effects on disguised face verification; an imperative task for surveillance settings.

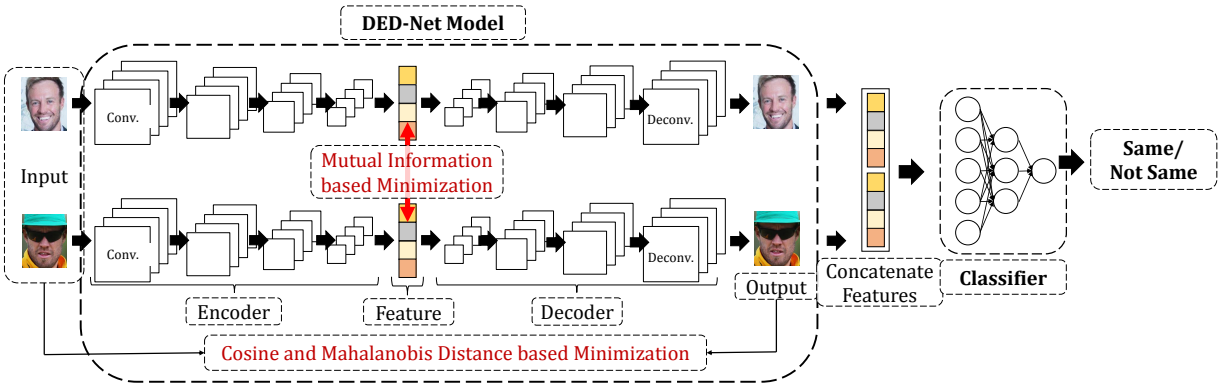



Figure 5-2: Overview of the proposed Disguise Resilient Framework on full face images. Pair of images are provided to the DED-Net model for feature extraction, followed by concatenation and input to the classifier. The classifier outputs a score denoting whether the two images belong to the same subject or not. The DED-Net is a convolutional encoder-decoder model, containing convolution (conv.) and deconvolution (deconv.) filters at different layers (represented by squares). It is optimized via the Cosine and Mahalanobis distance based minimization between the input and the reconstruction, along with the Mutual Information based loss between the features. The classifier is a classical neural network containing neurons in each layer (represented as circles). The different colors in the extracted feature signify different values at each position. Figure has been taken from the published manuscript [140].

## 5.3 Proposed Disguise Resilient (D-Res) Framework

Disguised face verification refers to the task of matching a given pair of facial images and classifying them as *genuine* (same class) or *imposter* (different class). Here, at least one of the images is disguised in nature i.e., it contains variations due to disguise accessories. Owing to the presence of disguise accessories, often parts of the facial region are obfuscated or have a different appearance than their original self (e.g. due to make-up or plastic surgery). Disguised face recognition thus suffers from its own set of unique challenges. Specifically:

- There exists high variations between images of an individual caused due to make-up or obfuscation due to artifacts (high intra-class variations).
- In cases where individuals intentionally alter facial features to appear similar to another individual, it leads to higher similarity between subjects (low inter-class variations).
- Disguised faces further present the challenge of no definitive facial region being altered or no specific kind of artifact being used. As described previously, there are several types of disguises: make-up, or use of accessories such as beard, moustache, sunglasses, or scarves,

which increases the diversity of the problem, making face recognition further challenging as well as pertinent.

Therefore, an effective disguised face verification algorithm must be able to extract discriminative features from the given image pair, useful for accurate verification. Further, the algorithm must be invariant to minor variations in the input image (such as light make-up or slight obfuscations).

In order to handle the above highlighted challenges, we propose an Encoder-Decoder model, termed as the Disguise Encoder-Decoder network (DED-Net). An Encoder-Decoder model consists of an encoder network which learns a representative feature for the given input, followed by a decoder network used to reconstruct the input image from the learned feature. Traditionally, in an encoder-decoder formulation, the optimization function minimizes the error between the input and the reconstructed output using the Euclidean distance. Euclidean distance has shown to perform well with images of a similar distribution, however, we postulate that a multi-objective loss function with different distance metrics is more suited to learn disguise resilient features. Thus, we propose the DED-Net model, which is further extended to present the Disguise Resilient (D-Res) framework for face verification. The D-Res framework utilizes tessellation on the input facial images for obtaining both local and global features, thus being useful for disguised face verification. The following subsections elaborate upon the proposed model and the proposed framework.

### 5.3.1 Disguise Encoder-Decoder Network (DED-Net)

The proposed model is designed such that it is able to learn representations while encoding the (i) "direction" variations between the image vectors, i.e. the locally altered features due to make-up or illumination (ii) "distribution" of pixel values, i.e. learn features resilient to noise and disturbance due to obfuscation or additional artifacts, while incorporating (iii) "supervision" during feature learning. The proposed model is thus better suited to handle disguise-specific variations caused due to the challenging nature of the problem. The proposed DED-Net formulation incorporates two distance metrics: *Cosine* and *Mahalanobis* for learning features invariant to disguise, both in forms of obfuscation and impersonation. Both the distance metrics are more resilient to non-identically and non-independently distributed feature vectors. This enables the feature learning model to incorporate the direction and magnitude of the loss between the input and its reconstruction. It

is our belief that for disguised face recognition, the Cosine distance is a suitable metric since it focuses on modeling the distribution of the pixels in two given images, as opposed to their exact pixel values. Therefore, in scenarios of variations in the pixel intensity values due to make-up, the Cosine distance will be able to model the similarity better as compared to the Euclidean distance. Further, for disguised face images, we believe that the Mahalanobis distance is a well-suited metric for an Encoder-Decoder architecture since it focuses on minimizing the reconstruction error based on a set of selected weight vectors, thus learning a model invariant to minor manipulations in the images. Due to the increased variability of data, we also incorporate a pair-based supervision term in the formulation of the proposed encoder-decoder model to enforce learning features invariant to disguise. This is accomplished by introducing *Mutual Information (MI)* as a penalty term in the loss function. If the mutual information is high, the dependence between the two vectors is high, thus resulting in good classification accuracy. Applying Mutual Information between the learned representations ensures similar features for same-class samples; a distinctive property useful for face verification models. Figure 5-2 presents a diagrammatic representation of the DED-Net model for an input pair of images. Thus, for a pair of images $(\mathbf{X}_1, \mathbf{X}_2)$ and label $y$, the proposed DED-Net model is mathematically expressed as:

$$\begin{aligned}\mathcal{L}_{D-Res} = \underbrace{-||\mathbf{X}_1 \odot \mathcal{D}(\mathcal{E}(\mathbf{X}_1))||^2 - ||\mathbf{X}_2 \odot \mathcal{D}(\mathcal{E}(\mathbf{X}_2))||^2}_{\mathcal{L}_C} + \\ \underbrace{||\mathbf{X}_1 \oplus (\mathcal{D}(\mathcal{E}(\mathbf{X}_1)))||^2 + ||\mathbf{X}_2 \oplus (\mathcal{D}(\mathcal{E}(\mathbf{X}_2)))||^2 + \ \lambda_M \|\mathbf{M}\|_1}_{\mathcal{L}_M} \\ \underbrace{- y * MI(\mathcal{E}(\mathbf{X}_1), \mathcal{E}(\mathbf{X}_2)) + \ (1-y) * MI(\mathcal{E}(\mathbf{X}_1), \mathcal{E}(\mathbf{X}_2))}_{\mathcal{L}_{MI}} + \lambda_R R\end{aligned} \tag{5.1}$$

where, $\mathcal{E}(.)$, $\mathcal{D}(.)$, and $MI(.)$ refer to the Encoder network, Decoder network, and the Mutual Information function. $\odot$ and $\oplus$ refer to the Cosine and Mahalanobis operator, respectively. $\mathbf{M}$ refers to the Mahalanobis matrix, $\lambda_M$ refers to its respective weight constant, and $\lambda_R R$ is the regularization term on the network weights. The first four terms ensure learning of representative features for the input pair via the Cosine and Mahalanobis loss functions, while the next two terms incorporate supervision for learning discriminative features. The proposed DED-Net model thus utilizes a multi-objective loss function for learning features useful for disguise recognition. Each

component of the proposed model has been described in detail below:

**Cosine based Loss Function ($\mathcal{L}_C$):** Cosine similarity models the similarity between two vectors in terms of their direction variations. It calculates the similarity based on the relationship of the vector values in contrast to the absolute magnitude difference between the two. Therefore, it has extensively been used in subspace learning algorithms that attempt to find vectors that best represent the given set of classes and for classification tasks [165]. In order to learn disguise resilient features, cosine similarity enables learning features invariant to make-up or local alterations which are common in cases of impersonation or passive disguises. We propose to utilize the Cosine similarity between the input and the output of an Encoder-Decoder model. The Cosine similarity is given as:

$$Cos(a, b) = \|a \odot b\|^2 = \frac{a \cdot b}{\|a\|_2^2 \times \|b\|_2^2} \tag{5.2}$$

where, $Cos(a, b)$ represents the Cosine similarity between two vectors $a$ and $b$, and $\odot$ represents the Cosine similarity operator. Incorporating the Cosine similarity in an Encoder-Decoder model with an encoder network as $\mathcal{E}(.)$ and a decoder network as $\mathcal{D}(.)$, results in the following formulation:

$$\mathcal{L}_C = -||\mathbf{X} \odot \mathcal{D}(\mathcal{E}(\mathbf{X}))||^2 + \lambda_R^{'} R \tag{5.3}$$

where, $\mathbf{X}$ refers to the input sample, $\mathcal{E}(\mathbf{X})$ refers to the learned representation, and $\mathcal{D}(\mathcal{E}(\mathbf{X}))$ refers to the reconstructions obtained by the decoder. $R$ is the regularizer on the network weights, and $\lambda_R^{'}$ is the regularization weight. As opposed to the Euclidean distance based encoder-decoder, the above model does not attempt to replicate the pixel values of the input data at the reconstruction layer, rather it learns representations such that the relationship between the pixels at the reconstruction is similar to that at the input.

**Mahalanobis based Loss Function ($\mathcal{L}_M$):** We propose using the Mahalanobis distance to model the distribution of reconstructed sample's pixel values with respect to the input sample. Mahalanobis distance accounts for the variability in the data distribution and is a unit-less scale-invariant distance metric which is used to measure the distance between two given points. For two vectors $a$ and $b$, the Mahalanobis distance is given as:

$$Mah(a, b) = \|a \oplus b\|^2 = (a - b)^T \mathbf{M} (a - b) \tag{5.4}$$

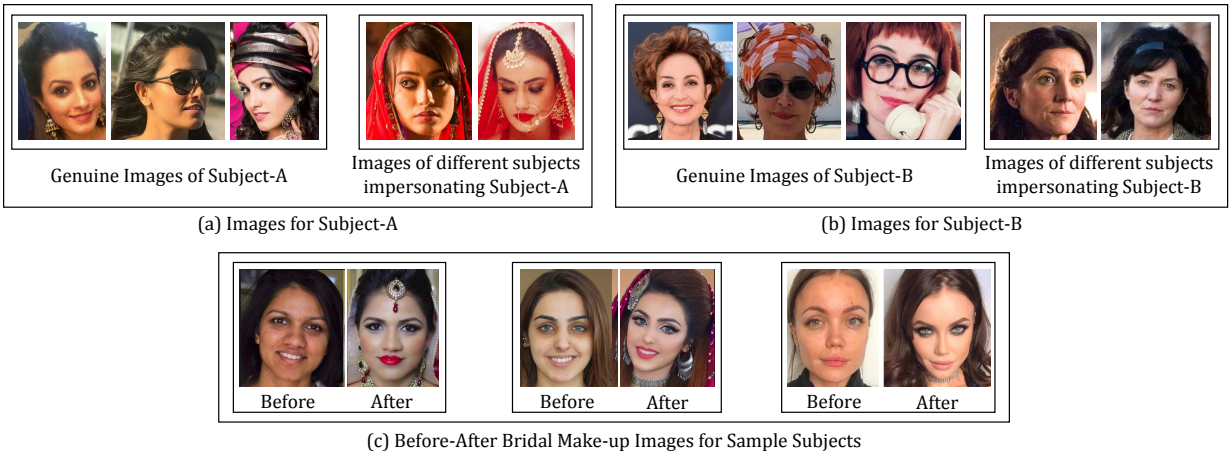



Figure 5-3: Sample images from the DFW2019 dataset having variations due to (a-b) disguise accessories demonstrating their genuine images (images belonging to the given subject (Subject-A or Subject-B)) and impersonator images (images of different subjects impersonating the given subject). The DFW dataset also contains images having variations due to (c) bridal make-up (sample images of three subjects have been shown before and after applying the make-up). The use of accessories results in obfuscated face regions, rendering automated face recognition challenging.

where, $Mah(a, b)$ represents the squared Mahalanobis (pseudo) distance between $a$ and $b$. $\mathbf{M}$ represents a pseudo-distance matrix having the dimensions $[m \times m]$, where $m$ corresponds to the vectorized dimension of the input sample. Traditionally, in Mahalanobis distance calculations, $\mathbf{M}$ is a symmetric positive semi-definite matrix, however, for minimizing the reconstruction error of the encoder-decoder model, these constraints are relaxed. The encoder-decoder formulation with Mahalanobis (pseudo) distance can be represented as:

$$\mathcal{L}_M = ||\mathbf{X} \oplus (\mathcal{D}(\mathcal{E}(\mathbf{X})))||^2 + \lambda_M \|\mathbf{M}\|_1 + \lambda_R'' R \tag{5.5}$$

Similar to Equation 5.3, $\lambda_R''$ enforces regularization on the network weights. In the above equation, we enforce a $\ell_1$-norm on the learned Mahalanobis matrix, that is, $\|\mathbf{M}\|_1$ with a regularization weight $\lambda_M$. $\ell_1$-norm promotes learning of a sparse $\mathbf{M}$ matrix, which forces the model to focus on some weight connections while minimizing the reconstruction loss. This enables the model to iteratively update its weights based on the reconstruction loss of a subset of the total nodes of the network. Therefore, minimizing the Mahalanobis distance ensures weight vectors are selected such that the distance between the input and its reconstruction is minimized when both are projected onto $\mathbf{M}$. This implies that the representation encodes information invariant to minor manipulation

of pixels and does not overfit on the training data.

**MI based Loss Function ($\mathcal{L}_{MI}$):** Adding the above mentioned distance metrics in the loss function ensures the learning of disguise resilient features. However, disguised faces bring with them the challenge of increased variations between images belonging to the same subject and also reduced differences between two different individuals due to similar accessories such as beard or scarf. Therefore, we add an additional supervision in the loss function via mutual information based penalty for a given pair of images. This results in learning discriminative features which enhances the recognition performance. Given two vectors $(a, b)$, Mutual Information (MI) is given as:

$$MI(a, b) = p(a, b)log\left(\frac{p(a, b)}{p(a)p(b)}\right) \tag{5.6}$$

We propose to incorporate mutual information between the learned features of an image pair as a penalty term to introduce supervision in the encoder-decoder model. Given two inputs, $\mathbf{X}_1$ and $\mathbf{X}_2$, mutual information is applied on the learned representations obtained by the Encoder ($\mathcal{E}(.)$). For a given pair of images and corresponding features, the MI based loss function is given as:

$$\mathcal{L}_{MI} = y * MI(\mathcal{E}(\mathbf{X}_1), \mathcal{E}(\mathbf{X}_2)) - \ (1 - y) * MI(\mathcal{E}(\mathbf{X}_1), \mathcal{E}(\mathbf{X}_2)) \tag{5.7}$$

where, $y$ is the label of the pair of the images i.e., whether they belong to the same subject ($y = 1$) or not ($y = 0$). Mutual information between the features of a pair of images is encoded as a supervised regularizer. Since mutual information is a similarity term, it is added in the loss function with a negative sign.

### 5.3.2 D-Res Framework for Face Verification with Disguise Variations

The proposed model has been used to prepare the D-Res framework for disguised face verification. Face verification models are required to process a pair of input images and output a score signifying whether the pair is *genuine* (belonging to the same class) or *imposter* (belonging to a different class). Figure 5-2 presents the D-Res framework for face verification using full face images. A pair of images are provided as input to the DED-Net model, resulting in a pair of feature vectors. The features are concatenated and provided as input to a three-dense layer neural network for clas-

sification into *same* or *different* class. The same pipeline is followed during the model training and testing.

**D-Res Framework on Tessellated Face Images:** Since disguised face recognition suffers from the challenge of arbitrary changes in different parts of the face image, the proposed D-Res framework also utilizes the tessellated face image. As has been observed in the literature [11, 16], tessellation of the face image enables an algorithm to focus on local features. Figure 5-1 presents a broad overview of the entire framework using the whole face and tessellated face as input. A pair of images are provided as input, followed by their tessellation into nine patches each. The generated patches are overlapping in nature, such that one-third of a given patch overlaps with the horizontal and vertical neighboring patches, respectively. Therefore, ten pairs consisting of the whole face and face patches are provided to the DED-Net model, followed by the classifier. For an input pair of images, there exist ten score vectors for *same* or *different* class (one for the full face pair, and nine for patch pairs). Weighted score-level fusion is performed on the scores to obtain the final score, as follows:

$$s = \sum_{i=1}^{10} w_i s_i \tag{5.8}$$

where, $w_i$ is the weight for the $i^{th}$ score ($s^i$). Final decision is taken based on the weighted score ($s$). At the time of testing, following Figure 5-1, the trained models are used for feature extraction and classification.

## 5.4 Experiments and Implementation Details

The effectiveness of the proposed D-Res framework has been evaluated on two datasets: (i) Disguised Faces in the Wild 2018 (DFW2018) dataset [143] and (ii) Disguised Faces in the Wild 2019 (DFW2019) dataset [136]. Figure 5-3 presents sample face images from the DFW2019 dataset. Experiments have been performed for the original high resolution images, where comparison has been drawn with the baseline results and the state-of-the-art results. Experiments have also been performed to present the baseline and benchmark results on low resolution images having resolutions $32 \times 32$, $24 \times 24$, and $16 \times 16$.

Table 5.1: Verification Accuracy (%) on the DFW2018 dataset. Owing to the same protocol, some results have directly been taken from the published paper [143]. The best performance is given in bold, while the second best has been underlined.

| **Algorithm** | **Protocol-1** | | **Protocol-2** | | **Protocol-3** | |
|---|---|---|---|---|---|---|
| | **1%FAR** | **0.1%FAR** | **1%FAR** | **0.1%FAR** | **1%FAR** | **0.1%FAR** |
| VGGFace [143] | 52.77 | 27.05 | 31.52 | 15.72 | 33.76 | 17.73 |
| ResNet-50 [143] | 73.94 | 38.48 | 54.86 | 31.55 | 56.22 | 32.68 |
| A-Link [153] | 95.73 | 75.38 | 88.97 | 72.13 | 89.30 | 72.72 |
| AEFRL [148] | 96.80 | 57.64 | 87.82 | 77.06 | 87.90 | 75.54 |
| DenseNet+COST [154] | 92.10 | <u>62.20</u> | 87.10 | 72.10 | 87.60 | 71.50 |
| MiRA-Face [187] | 95.46 | 51.09 | 90.65 | 80.56 | 90.62 | 79.26 |
| ArcFace [22] | **98.66** | 60.84 | <u>95.08</u> | **92.20** | <u>95.11</u> | <u>91.76</u> |
| FLF [84] | - | - | - | - | 91.30 | 78.55 |
| DDFR [199] | 96.30 | 60.84 | 90.19 | 80.61 | 90.30 | 79.57 |
| TeCS$^2$ + DenseNet [155] | 96.9 | 65.3 | 90.6 | 79.2 | 90.9 | 79.8 |
| **Proposed** | <u>98.02</u> | **93.16** | **96.83** | <u>90.58</u> | **97.18** | **91.80** |

**DFW2018 Dataset** [143]: The DFW2018 dataset is a CVPR2018 competition dataset, containing 11,157 face images of 1,000 subjects. The dataset contains face images collected from the Web, demonstrating variations across different disguise accessories such as scarves, hats, sunglasses, turbans, and beards. For a given subject, the dataset contains two non-disguised face images, multiple disguised images, and multiple impersonator images. The impersonators of a subject correspond to different people who appear to impersonate or look like them. The dataset contains three pre-defined protocols: (i) Impersonation, (ii) Obfuscation, and (iii) Overall. Protocol-1 focuses on evaluating a face recognition algorithm for impersonators, that is, pairs of images which appear to belong to the same subject but do not, while the second protocol focuses on evaluating an algorithm under disguise variations for genuine pairs. The third protocol evaluates an algorithm on the entire dataset. Experiments have been performed on all three protocols: pre-defined training testing split is followed, where images pertaining to 400 subjects form the training partition (3,386 images) while the remaining 600 subjects form the test set (7,771 images).

**DFW2019 Dataset** [136]: The DFW2019 dataset is an ICCV2019 competition dataset, containing 3,840 face images of 600 subjects. The dataset encompasses variations across disguise acces-

sories, bridal make-up, and plastic surgery. Since it has been collected from the Web, the dataset contains unconstrained images with variations due to pose, illumination, acquisition device, gender, and ethnicity. Four protocols have been defined on the DFW2019 dataset: (i) Impersonation, (ii) Obfuscation, (iii) Plastic Surgery, and (iv) Overall. Protocols 1,2,4 are similar to those of the DFW2018 dataset, whereas Protocol-3 focuses on evaluating an algorithm under variations observed due to plastic surgery. Experiments have been performed on all protocols, and consistent with the pre-defined protocol, all images have been used as the test set, and no explicit training has been performed on the DFW2019 dataset, thus resulting in a cross-dataset evaluation protocol.

**Implementation Details:** Figure 5-2 presents the complete pipeline for the proposed D-Res framework. For all the experiments, a pre-trained ResNet-50 [47] based Encoder-Decoder model is used as the base architecture, followed by a three dense layer neural network as the classifier. The Mahalanobis matrix (M) is initialized as an identity matrix, and updated using gradient descent during model training. Training has been performed using a Nvidia K40 GPU, with an initial learning rate of $1e-4$ using the Adam optimizer [69] for 200 epochs, and a batch-size of 50. The weights for Eq. 5.8 and the $\lambda$ values (Eq. 5.1) have been obtained empirically via grid search [9]. The value of the $\lambda$ parameters varies in the range of $0.01-0.1$ for different experiments (Eq. 5.1), and Dropout [53] has been used as the regularizer ($R$). Specifically, for the DFW2018 and 2019 datasets, $\lambda_R = 0.1$ and $\lambda_M = 0.02$. Data augmentation has been performed during training by means of flipping along the y-axis, color variations, and adding Gaussian noise. Bicubic interpolation has been used for increasing/decreasing the resolution of the face images for different experiments.

## 5.5 Results and Analysis

Tables 5.1-5.4 and Figures 5-6-5-7 present the results obtained on the DFW2018 and DFW2019 datasets. Pre-defined metrics have been used for reporting the results: GAR@1% FAR and GAR@0.1% FAR for the DFW2018 dataset, while GAR@0.1% FAR and GAR@0.01% FAR have been used with the DFW2019 dataset. The performance of the proposed D-Res framework along with the baseline results have also been provided for varying resolutions of the input image, that is, $32 \times 32$, $24 \times 24$, and $16 \times 16$. Beyond disguised face verification, additional results have

also been reported on standard benchmark face verification datasets.

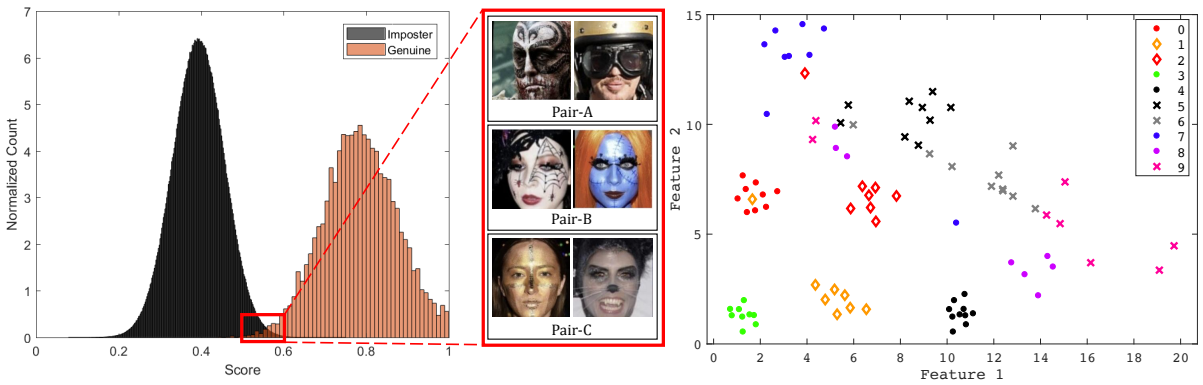
Table 5.2: Genuine Acceptance Rate (GAR) (%) on the DFW2019 dataset for Protocol-1 (Impersonation), Protocol-2 (Obfuscation), Protocol-3 (Plastic Surgery), and Protocol-4 (Overall) for two False Acceptance Rates: 0.1% and 0.01%. Comparative results have directly been taken from the published manuscript [136]. The best performance is given in bold, while the second best has been underlined.

| **Algorithm** | **Protocol-1** | | **Protocol-2** | | **Protocol-3** | | **Protocol-4** | |
|---|---|---|---|---|---|---|---|---|
| | **0.1%** | **0.01%** | **0.1%** | **0.01%** | **0.1%** | **0.01%** | **0.1%** | **0.01%** |
| Baseline (ResNet-50) | 47.6 | 38.4 | 35.3 | 16.4 | 46.4 | 22.4 | 35.9 | 16.8 |
| Baseline (LightCNN) | 74.4 | <u>51.2</u> | 55.5 | 36.9 | 69.2 | 47.2 | 55.7 | 36.5 |
| XuXu | 66.0 | 24.4 | 90.5 | 80.5 | 90.0 | 81.2 | 90.0 | 76.0 |
| LightCNNDFW | 70.4 | 43.2 | 57.5 | 38.6 | 70.8 | 43.6 | 57.1 | 37.4 |
| Composite Mini Batch [149] | 52.4 | 1.2 | 92.3 | 87.7 | 95.6 | 92.0 | 92.1 | 83.1 |
| FEBNet [151] | 54.8 | 42.4 | 92.3 | 87.6 | 78.8 | 47.6 | 90.7 | 73.6 |
| ArcFaceInter [22] | <u>72.4</u> | 44.8 | 98.7 | <u>97.9</u> | **98.4** | <u>95.6</u> | 98.3 | 92.0 |
| ArcFaceIntraInter [22] | 56.8 | 17.6 | <u>98.9</u> | **98.4** | **98.4** | <u>95.6</u> | <u>98.4</u> | <u>93.6</u> |
| **Proposed** | **79.2** | **55.2** | **99.2** | 97.3 | <u>98.0</u> | **96.0** | **98.7** | **96.3** |

### 5.5.1 Performance of the D-Res Framework

**Comparison with State-of-the-art Algorithms on the DFW2018 Dataset:** Table 5.1 presents the performance of the D-Res framework, along with the other comparative algorithms and baseline results on the DFW2018 dataset. Owing to the same protocol, results have directly been taken from the published manuscripts. The effectiveness of the proposed framework can be observed across all three protocols, where it achieves the state-of-the-art or second best performance. For example, on the Impersonation protocol, the proposed framework demonstrates a substantial increase at 0.1%FAR, by achieving 93.16%, showcasing an improvement of over 18% as compared to the current state-of-the-art (A-Link: 75.38%). Similar results can be observed for the other protocols at other FARs as well. Overall, the proposed framework achieves over 96% for all protocols at 1%FAR, while obtaining over 90% for all protocols at 0.1%FAR. The substantial improvement observed in the challenging Impersonation protocol strengthens the utility of the proposed framework for disguised face verification.

**Comparison with State-of-the-art Algorithms on the DFW2019 Dataset:** Similarly, Table 5.2




(a) Score Histogram (b) tSNE Visualization

Figure 5-4: (a) Histogram of the scores obtained by the D-Res framework on the Overall protocol (DFW2019). Separation can be observed between the genuine and imposter scores with a small overlap. Sample genuine pairs (images belonging to the same subject) are also shown which were not identified by the D-Res framework. The pairs had scores in the overlapping region between the genuine and imposter scores. Excessive make-up and obfuscation results in highly challenging samples. (b) tSNE visualization of the features learned by the D-Res framework for ten subjects suggesting distinguishing features based on the subject information.

presents the performance of the D-Res framework along with other comparative techniques and baseline results on the DFW2019 dataset. The proposed framework achieves comparative performance to the state-of-the-art results, often resulting in improved performance as well. For example, at the Overall protocol (Table 5.2), the proposed framework presents an improvement of at most 3% at both FARs as compared to the state-of-the-art algorithm (ArcFaceIntraInter). Similar to the previous dataset, the proposed framework demonstrates a substantial improvement on the Impersonation protocol, especially at a lower FAR. An improvement of around 4% is observed at 0.01% FAR (55.2% from 51.2%). The improvement obtained by the proposed framework, especially at lower FARs, promotes its usage for real world scenarios requiring high performance at stricter FARs. Figure 5-4(a) presents the score distribution obtained by the D-Res framework demonstrating minimal overlap between the scores, thus suggesting high discriminability between the features. Figure 5-4(a) also presents sample mis-classifications of the proposed D-Res framework. Make-up resulting in obfuscation of true facial features also makes automated recognition challenging. Further, Figure 5-4(b) presents the tSNE visualization of the features obtained from the DED-Net model. The plot demonstrates distinction between different subjects, possibly due to the Mutual Information loss component which ensures samples of the same class have similar

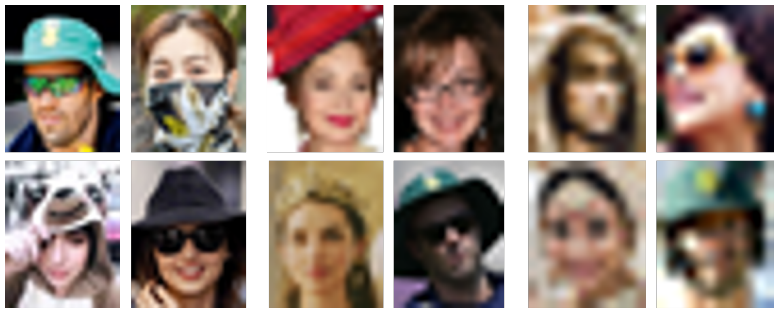
Table 5.3: Ablation study of the D-Res framework on the overall protocol of the DFW2018 (Protocol-3) and DFW2019 (Protocol-4) datasets. GAR at 0.1% FAR has been reported, and the efficacy of each component in the framework is evaluated by computing the performance of the framework without the component.

| **Algorithm** | **DFW2018** | **DFW2019** |
|---|---|---|
| Proposed - {Cosine} | 85.6 | 94.0 |
| Proposed - {Mahalanobis} | 86.4 | 94.5 |
| Proposed - {MI} | 84.0 | 91.9 |
| Proposed - {Tessellation} | 78.5 | 90.4 |
| **Proposed D-Res Framework** | **91.8** | **98.7** |

representations, while samples of different classes have varying representations. The class-specific features learned by the DED-Net model further support the D-Res framework during the verification process, resulting in distinguishable features for genuine/imposter prediction.

**Ablation Study on the D-Res Framework:** Table 5.3 presents the verification accuracy is reported at 0.1% for the DFW2018 and DFW2019 datasets, on the Overall protocols. The ablation study is performed by removing a single component from D-Res framework while keeping the remaining as it is. Maximum drop of around 13% and 8% (on DFW2018 and DFW2019, respectively) is observed upon removing the component of tessellation ('Proposed - Tessellation'). The accuracy drop reinstates the need for incorporating both local and global features for disguised face verification. The second major contribution comes from the component of Mutual Information (MI), where a drop of around 7% and 4% (on the DFW2018 and DFW2019 dataset, respectively) is observed upon its removal from the proposed framework ('Proposed - MI'). Mutual information enforces similar features for different images of the same subject, thereby modeling the intra-class variations. Removal of the MI based component might result in less discriminative features, thus resulting in a drop in performance. Similarly, Table 5.3 presents the variations observed upon removing the Cosine/Mahalanobis based component. The drop in performance due to the removal of each component strengthens their inclusion in the D-Res framework.

For a given pair of input images, the D-Res framework takes around 0.02 seconds for feature extraction and classification (genuine/imposter) on a V100 GPU. During inference, the proposed framework utilizes a lightweight ResNet-50 base architecture for feature extraction, which



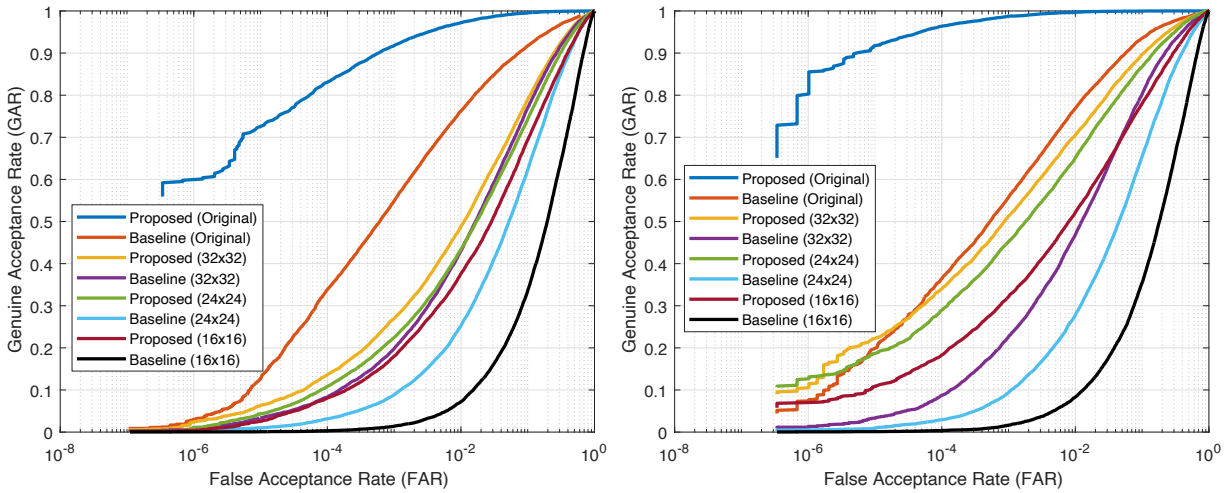
(i) 32x32 (ii) 24x24 (iii) 16x16

Figure 5-5: Images from the DFW2019 dataset at different resolutions. Images have been bicubically interpolated to $224 \times 224$.

is similar to or smaller than other architectures used in recent literature (ArcFace [20]: ResNet-50/ResNet-100, GroupFace [68]: ResNet-100). Over the base architecture, the proposed D-Res framework also performs tessellation followed by score fusion for obtaining the final output. A marginal increase of 0.01 seconds is obtained on incorporating the tessellation step which results in an overall performance improvement of over 8% (obtained via ablation study in Table 5.3).

### 5.5.2 Baseline Results and Performance of D-Res Framework for Low Resolution Disguised Face Verification

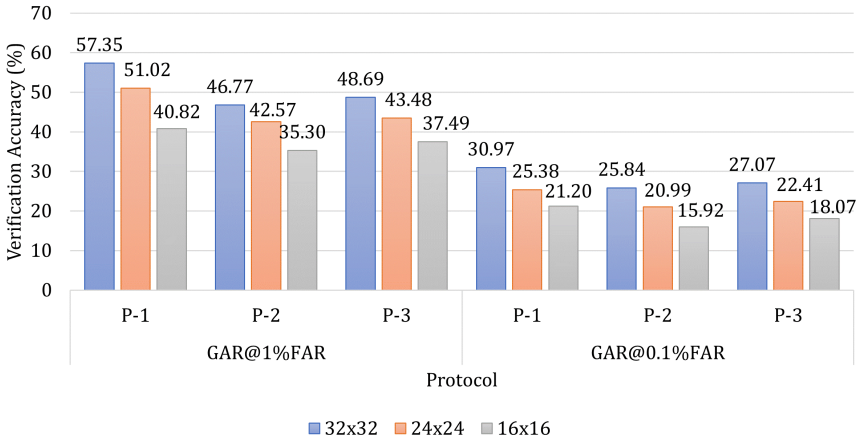
As discussed previously, disguised face recognition has wide-spread applicability in surveillance scenarios, where low resolution data is often captured via CCTV cameras. Low resolution face recognition has attracted the attention of the research community since the past few decades [61, 110, 171], with research focusing either on super-resolution/synthesis techniques [182, 183], learning resolution invariant features/classifiers [37, 88, 138]. However, to the best of our knowledge, no research has focused on understanding or evaluating low resolution disguised face recognition. To this effect, as part of this research, we present the baseline results for low resolution disguised face recognition, along with the results of the proposed framework. Evaluation has been performed on three resolutions: $32 \times 32$, $24 \times 24$, and $16 \times 16$. Figure 5-6(a) presents sample images at different resolutions, demonstrating the decreasing information content with reducing resolutions. Baseline results have also been computed using the pre-trained LightCNN-29 model [177], which was also used as the baseline for DFW2019 competition (Table 5.2).



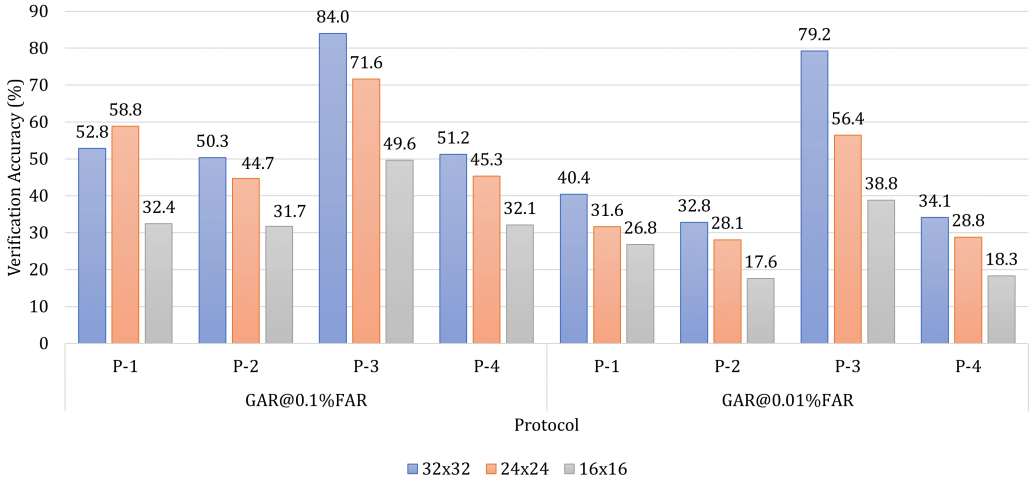

(a) DFW2018 dataset (b) DFW2019 dataset

Figure 5-6: ROC curves on the (a) DFW2018 and the (b) DFW2019 datasets, for multiple resolutions at the Overall protocol. The performance of the D-Res framework has been compared with the baseline model (LightCNN-29) for different resolutions: (i) Original, (ii) 32x32, (iii) 24x24, and (iv) 16x16. The D-Res framework outperforms the respective baseline at each resolution.

Figure 5-6(b-c) presents the Receiver Operating Characteristic (ROC) curves on the Overall protocol for the D-Res framework and the baseline model for different resolutions. For both the datasets, the proposed framework outperforms the corresponding baseline performance by a respectable margin. For example, on the DFW2018 dataset, the proposed framework achieves 43.48% at 1% FAR for $24 \times 24$ resolution, whereas the baseline model achieves 25.28%. Similar improvement is seen for the DFW2019 dataset as well, across different resolutions. Figure 5-7 presents the performance of the D-Res framework on all seven protocols, for the three different resolutions, at the pre-defined FARs for both the datasets. While it is not surprising to observe a drop in performance as the resolution decreases, a substantial drop in performance is observed between the performance on $32 \times 32$ and $16 \times 16$ face images. For example, on the DFW2019 dataset, at 0.1% FAR, the proposed framework obtains an accuracy of 84.0% at $32 \times 32$, which reduces to 49.6% at $16 \times 16$. While the D-Res framework achieves improved performance as compared to the baseline, lower resolutions still present vast scope for improvement. In order to deploy face recognition systems in real world scenarios, we believe it is imperative for future research to focus on the challenging task of disguised face recognition with low resolution variations.




(a) DFW2018 dataset




(b) DFW2019 dataset

Figure 5-7: Bar graphs demonstrating the variation in accuracy obtained over the DFW2018 and DFW2019 datasets for 32x32, 24x24, and 16x16 input resolutions.

### 5.5.3 Additional Results: Benchmark Face Verification Datasets

The proposed DED-Net model has also been evaluated on traditional benchmark face verification datasets. Evaluation is performed on three standard datasets:

- Labeled Faces in the Wild (LFW) dataset [57] consists of 13,233 unconstrained facial images belonging to 5,749 identities. The standard *unrestricted with labelled outside data* protocol has been followed, which contains 6,000 image pairs for face verification evaluation.

- Youtube Face (YTF) dataset [175] consists of 3,425 videos of 1,595 identities. The face verification protocol requires matching 5,000 video pairs in 10 folds and report the aver-

Table 5.4: Verification accuracy (%) of the DED-Net model and comparative techniques on the LFW, YTF, and IJB-B datasets. Comparative results have directly been taken from the different published manuscripts.

| **Algorithm** | **LFW** | **YTF** | **IJB-B** | |
|---|---|---|---|---|
| | | | **FAR=1e-4** | **FAR=1e-5** |
| RegularFace [192] | 99.61 | 96.7 | - | - |
| CosFace [165] | 99.81 | 97.6 | 94.80 | 88.11 |
| UniformFace [29] | 99.80 | 97.7 | - | - |
| AFRN [65] | 99.85 | 97.1 | 88.50 | 77.10 |
| ArcFace [20] | 99.83 | 97.7 | 94.25 | 89.33 |
| GroupFace [68] | 99.85 | 97.8 | **94.93** | 91.24 |
| Circle Loss [152] | 99.73 | 96.3 | - | - |
| TeCS$^2$ + DenseNet [155] | 99.40 | - | - | - |
| SFace [195] | 99.83 | 98.0 | - | - |
| Orthogonality Loss [180] | 99.7 | - | 94.33 | - |
| **Proposed** | **99.87** | **98.2** | **94.89** | **91.26** |

age accuracy. Similar to the LFW dataset, results have been reported using the standard *unrestricted with labelled outside data* protocol.

- IARPA Janus Benchmark-B Face (IJB-B) dataset [174] contains 21.8K still images (facial and non-facial), 55K frames from 7,011 videos corresponding to 1,845 subjects. Results have been reported on the standard pre-defined face verification protocol involving identifying a given pair of images as genuine or imposter. Face verification accuracy has been reported for the False Acceptance Rate (FAR) of $1e-4$ and $1e-5$.

Table 5.4 presents the results obtained by the proposed face verification framework, along with the recent comparative techniques. Comparison has been performed with recent angular-margin based Softmax losses (such as CosFace [165] and ArcFace [20]), domain-specific GroupFace [68], along with other recent techniques such as RegularFace [192] and UniformFace [29]. As shown in Table 5.4, the proposed framework outperforms existing techniques on the LFW and YTF datasets, by obtaining 99.87% and 98.23%, respectively. Similar performance is obtained on the IJB-B dataset, where the proposed framework achieves 94.89% and 91.26% accuracy for $FAR = 1e-4$ and $FAR = 1e-5$, respectively. The proposed framework is thus amongst the top two performing techniques for the IJB-B dataset. The performance obtained on the challenging benchmark face verification datasets further strengthens the utility of the proposed framework for generic facial

verification tasks as well.

## 5.6 Summary

Automated disguised face recognition is a long-standing problem and its applicability in scenarios such as access control and surveillance suggests high real world utility. To this effect, this research proposes a novel Disguise Resilient (D-Res) framework for effective disguised face verification. The proposed framework utilizes the proposed Disguise Encoder-Decoder network (DED-Net) formulation for extracting meaningful disguise-invariant features. The DED-Net model is optimized using a multi-objective loss function which models the Cosine and Mahalanobis distance between the input and the reconstructions, while introducing Mutual Information based discrimination at the feature level. The effectiveness of the proposed framework is demonstrated on two recent and challenging benchmark datasets: (i) DFW2018 and (ii) DFW2019 datasets, where it achieves state-of-the-art performance in almost all seven protocols, with substantial improvement specifically for the Impersonator protocols. This research also presents the baseline results and the performance of the D-Res framework for low resolution disguised face recognition. Results have been demonstrated on three resolutions: $32 \times 32$, $24 \times 24$, and $16 \times 16$. A sharp drop in performance is observed from $32 \times 32$ to $16 \times 16$ resolution, thus suggesting a need for dedicated research focus. Despite the challenging nature of the problem, the D-Res framework demonstrates improvement as compared to the baseline performance, however, the lower accuracies demand dedicated research attention, especially for low resolution disguised face verification.

Beyond disguise face verification, the proposed framework has also been evaluated on standard benchmark face verification datasets, where it achieves improved performance, thus suggesting applicability in generic face verification scenarios as well. Further, the ablation study performed on the framework suggests contribution of tessellation and Mutual Information based components; both of which can be incorporated in traditional classification algorithms for potential boost in performance. Mutual information based loss promotes learning of discriminative features, while tessellating the input image into patches enables the model to focus on different local and global features. In future, the D-Res framework can be optimized to reduce the additional cost while maintaining the high recognition performance. The framework achieves improved performance

across datasets and protocols; however, dedicated effort is required for further enhancing the face verification performance for scenarios of impersonation (79.2% obtained on Protocol-1, DFW2019 dataset as compared to 98.7% on the overall dataset). This research has been published in IEEE Transactions on Circuits and Systems for Video Technology (IEEE T-CSVT). All images of this chapter have been taken from the published manuscript [140].

THIS PAGE INTENTIONALLY LEFT BLANK

# Chapter 6

# Gender Prediction from Very Low Resolution Face Images

## 6.1 Introduction

Gender is one of the primary attributes often used to describe an individual. Over the past several decades, researchers have attempted to understand the behavioral differences between the two genders [31, 89]. These behavioral difference have further been explored for enhancing digital user experience, human computer interaction, and gender targeted advertisements. On the other hand, human beings have also always used the physical differences as a key identifying attribute of an individual. This has resulted in utilization of gender information in scenarios of surveillance, security, and access control. Given the tremendous applications of automated gender classification, researchers have proposed several novel algorithms using different biometric modalities to model the same [39, 87, 106, 124, 159].

Face of an individual is one of the most distinguishable and non-invasive modalities for gender recognition. Automated gender recognition[1] on face images has attracted the attention of researchers since a long time. While gender recognition in controlled, well-illuminated scenarios has been well explored, it is still considered an arduous task in unconstrained scenarios. For example, low resolution faces captured in uncontrolled surveillance settings are difficult to process by most of the existing automatic gender classification algorithms. In general, images captured

[1] Gender classification and gender recognition have been used interchangeably.

Figure 6-1: Sample male images from SCface dataset [42] captured from surveillance cameras.

from surveillance cameras or from a distance, entail non-cooperative subjects in unconstrained environments. These images are often of poor quality, resulting in low resolution face regions. From Figure 6-1, it can be observed that low resolution face images often contain less information content along with several challenging covariates. These challenges require developing a robust algorithm which is capable of performing gender classification in the given challenging scenarios.

In the literature, researchers have explored several techniques to address the task of gender recognition. Moghaddam *et al.* [99] presented the superior performance of non-linear Support Vector Machines for performing gender recognition in low-resolution thumbnail images. Comparative analysis with other techniques such as Fisher linear discriminant and nearest neighbor classifiers presented the strength of their proposed approach. Andreu *et al.* [6] varied the resolution of face images from $2 \times 1$ to $329 \times 264$ pixels and studied it's effect on gender recognition for large datasets. Experiments were performed on pixel intensity values with two classifiers, and reduced performance was observed with low resolution faces images. In 2016, Juefei-Xu *et al.* [63] proposed DeepGender, a progressive convolutional neural network training technique for gender recognition, with an application to low resolution face images. The model aimed to learn important regions of the face for the given task and yielded promising results.

This research aims to address the task of gender classification in (very) low resolution images. Two supervised autoencoder formulations are proposed for learning discriminative features, useful for effective classification. Specifically, a *Class Specific Mean Autoencoder* is proposed, which uses the class information of a given sample at the time of training to learn the intra-class similarity and extract similar features for samples belonging to the same class. Further, a novel formulation of *Class Representative Autoencoder* is also presented which encodes both inter-class and intra-class features for feature learning. For the specific task of gender classification, *Autogen* is utilized, which is a Class Representative Autoencoder model for gender classification on

(a) Male (b) Female

Figure 6-2: Mean-male and mean-female images obtained from the CMU Multi-PIE dataset [44].

low resolution face images. During the feature learning process, AutoGen aims to incorporate the inherent characteristics of male and female facial features.

## 6.2 Proposed Supervised Autoencoders

Deep learning architectures have been used in literature to address a large variety of tasks [77]. Specifically, recent models such as the FaceNet [134], VGG-Face [113], and DeepFace [157] have shown high performance for the task of face recognition. Models have been developed to perform automated face detection and alignment as well [17, 30, 80]. Figure 6-2 presents the mean images of the two classes, demonstrating significant variation. Based on this observation, it is our hypothesis that projecting the image features closer to the class mean can assist in learning class specific discriminative features. Therefore, in this work, we propose two models: (i) Class Specific Mean Autoencoder and (ii) Class Representative Autoencoder, which learn features while modeling the class variations, such that the learned features are discriminative in nature. Before elaborating upon the proposed models, the following subsection presents some preliminaries.

### 6.2.1 Preliminaries: Supervised Autoencoders

Several researchers have proposed modifications to the traditional autoencoder architecture. Table 6.1 provides a summary of these architectures. Most of these are unsupervised in nature, however, researchers have proposed supervised architectures that leverage the availability of labeled data as well. In this section, we briefly present the original formulation of autoencoder followed by discussing the existing supervised architectures.

Table 6.1: Brief literature review of autoencoder based formulations.

| Authors | Approach | Supervised |
|---|---|---|
| Vincent *et al.* [162] | Stacked Denoising Autoencoder (SDAE): Noise is added to the input data to learn robust representations. | No |
| Ng [105] | Incorporated $\ell_1$ norm in the loss function of the autoencoder to introduce sparsity in the learned features. | No |
| Rifai *et al.* [128] | Contractive Autoencoder (CAE): Input space is localised by adding a penalty term which is the Jacobian of the input with respect to the hidden layer. | No |
| Rifai *et al.* [127] | Higher order Contractive autoencoder: CAE + Hessian of the output with respect to the input. | No |
| Zheng *et al.* [194] | Contrastive autoencoder: A term to reduce the intra-class variations between the learned representation of samples belonging to the same class is added at the final layer. | Yes |
| Wang *et al.* [168] | Generalised Autoencoder: SDAE is modified such that the representation incorporates the structure of the dataspace. | No |
| Zhang *et al.* [191] | Stacked Multichannel Autoencoder: The gap between synthetic and real data is reduced by learning a mapping between the two. | No |
| Gao *et al.* [34] | Inspired from SDAE, an identification model is proposed, where the probe image is treated as the noisy input while the gallery images are treated as the clean input. | Yes |
| Zhuang *et al.* [200] | A two layer model is proposed wherein, a representation is learned in the first layer, and the class label is encoded in the second layer. | Yes |
| Majumdar *et al.* [90] | A joint sparsity (using $\ell_{2,1}$) promoting supervision penalty term is added to the loss function of SDAE. | Yes |
| Meng *et al.* [97] | A relational term, which aims to model the relationship between the input data is added to the loss function. | No |

For a given input $x$, the loss function of a single layer traditional autoencoder [49] is given as follows:

$$\underset{\mathbf{W}_e, \mathbf{W}_d}{\operatorname{argmin}} \|x - \mathbf{W}_d \phi(\mathbf{W}_e x)\|_2^2 \tag{6.1}$$

where, $\mathbf{W}_e$ and $\mathbf{W}_d$ are the respective encoding and decoding weights of the autoencoder, and $\phi$ corresponds to an activation function, generally incorporated for introducing non-linearity in the model. Common examples of activation functions are $sigmoid$ and $tanh$. An autoencoder learns features ($f_x = \phi(\mathbf{W}_e x)$) of the given input $x$, such that the error between the original sample and it's reconstruction ($\mathbf{W}_d f_x$) is minimized. For a $k$ layered autoencoder, having encoding weights as $\mathbf{W}_e^1, \mathbf{W}_e^2, ..., \mathbf{W}_e^k$, and decoding weights as $\mathbf{W}_d^1, \mathbf{W}_d^2, ..., \mathbf{W}_d^k$, the loss function of Equation 6.1 is

modified as follows:

$$\underset{\mathbf{w}_e^1,\ldots,\mathbf{w}_e^k,\mathbf{w}_d^1,\ldots,\mathbf{w}_d^k}{\operatorname{argmin}} \|x - b \circ a\,(x)\|_2^2 \tag{6.2}$$

where, $a(x) = \phi(\mathbf{W}_e^k(\phi(\mathbf{W}_e^{k-1}\ldots(\phi(\mathbf{W}_e^1 x)))))$ refers to the encoding function, and $b\,(x) = \mathbf{W}_d^1(\mathbf{W}_d^2\ldots(\mathbf{W}_d^k x))$ corresponds to the decoding function. The first and the last layers correspond to the input and output layers respectively, while the remaining layers are often termed as the hidden layers.

In the literature, researchers have incorporated class information in the traditional formulation of an autoencoder in order to facilitate supervision. Gao *et al.* [34] modify the denoising autoencoder [162] to learn supervised image representations in order to optimize the identification performance. At the time of training, for a given subject, the probe image is the input to the autoencoder (analogous to the noisy input), and the gallery image of the subject (analogous to the clean image) is the target image used for computing the reconstruction error, as in the case of a denoising autoencoder. A similarity preservation term is added to the loss function such that the samples belonging to the same class have a similar representation. Given probe and gallery images of class $i$, each probe image is represented using $x_{ni}$ and its corresponding gallery images are represented using $x_i$. The loss function for the supervised autoencoder is as follows:

$$\begin{aligned}\underset{\substack{\mathbf{w}_e^1,\ldots,\mathbf{w}_e^k,\\ \mathbf{w}_d^1,\ldots,\mathbf{w}_d^k}}{\operatorname{argmin}} \frac{1}{N}\sum_i \Big( \|x_i - b \circ a(x_{ni})\|_2^2 + \lambda \|a(x_i) - a(x_{ni})\|_2^2 \Big) \\ +\alpha\Big(KL(\rho_x\|\rho_o) + KL(\rho_{x_n}\|\rho_o)\Big) \\ where, \rho_x = \frac{1}{N}\sum_i \frac{1}{2}\Big(a(x_i)+1\Big) \quad and\ \rho_{x_{ni}} = \frac{1}{N}\sum_i \frac{1}{2}\Big(a(x_{ni})+1\Big)\end{aligned} \tag{6.3}$$

here, the first term corresponds to the reconstruction error, second is the similarity preservation term, and the remaining two terms correspond to the Kullback Leibler (KL) divergence [71] to introduce sparsity in the hidden layers.

Contrastive Autoencoder (CsAE) proposed by Zheng *et al.* [194], is another variant of supervised autoencoder which uses the class label information during training. The loss function of the model is the difference between the output of two sub-autoencoders trained simultaneously on samples belonging to the same class, along with the loss function of each sub-autoencoder. The

equation for the same is given as:

$$\underset{\substack{\mathbf{W}_e^1,\ldots,\mathbf{W}_e^k,\\ \mathbf{W}_d^1,\ldots,\mathbf{W}_d^k}}{\operatorname{argmin}} \lambda(\|x_1 - b \circ a(x_1)\|_2^2 + \|x_2 - b \circ a(x_2)\|_2^2) \; + (1-\lambda)\, \|O_k(x_1) - O_k(x_2)\|_2^2 \,. \tag{6.4}$$

where, $x_1$ and $x_2$ represent two different input samples belonging to the same class. For each sub-autoencoder, $a(x) = \phi(\mathbf{W}_e^k \phi(\mathbf{W}_e^{k-1} ... \phi(\mathbf{W}_e^1(x)))$ and $b(x) = \mathbf{W}_d^1(\mathbf{W}_d^2 ... \mathbf{W}_d^k(x))$, where $\mathbf{W}_e^i$ and $\mathbf{W}_d^i$ refer to the encoding and decoding weights of the $i^{th}$ layer, and $O_k(x)$ is the output of the $k^{th}$ layer.

Recently, Majumdar *et al.* [90] present a class sparsity based supervised encoding algorithm wherein a joint-sparsity promoting $l_{2,1}-$norm supervision penalty is added to the loss function. For samples $\mathbf{X}$, belonging to total $C$ classes, the modified algorithm is presented as:

$$\underset{\mathbf{W}_e^1,\ldots,\mathbf{W}_e^k,\mathbf{W}_d^1,\ldots,\mathbf{W}_d^k}{\operatorname{argmin}} \|\mathbf{X} - b \circ a(\mathbf{X})\|_F^2 + \lambda \sum_{c=1}^{C} \|\mathbf{W}_e \mathbf{X}_c\|_{2,1} \tag{6.5}$$

where, $\mathbf{X}_c$ refers to the samples belonging to class $c$. The regularization term enforces same sparsity signature across each class, which leads to similar representations of samples from a given class.

### 6.2.2 Proposed Class Specific Mean Autoencoder

While all the above techniques incorporate supervision into an otherwise unsupervised model, the proposed architecture incorporates the mean feature of each class into the feature learning process as well. The key motivation behind the proposed algorithm lies in the observation that a sample belonging to class *male* is closer to the mean image of class *male* as compared to the mean image of class *female*. Thus it is our hypothesis that if the intra-class variations are encoded, it may help in learning class-specific features. Inspired from this observation, in this research, we present a novel formulation of Class Specific Mean Autoencoder.

In the proposed formulation, the loss function of an autoencoder [49] is updated by introducing

class specific information. For simplicity and clarity, Equation 6.1 is repeated as follows:

$$\underset{\mathbf{W}_e, \mathbf{W}_d}{\operatorname{argmin}} \|x - \mathbf{W}_d \phi(\mathbf{W}_e x)\|_2^2 \tag{6.6}$$

For an input sample $x_c$, belonging to class $c$, the feature vector $f_{x_c}$ is defined as follows:

$$f_{x_c} = \phi(\mathbf{W}_e x_c) \tag{6.7}$$

The mean feature vector pertaining to the $c^{th}$ class is defined as:

$$m_c = \mu(\phi(\mathbf{W}_e \mathbf{X}_c)) \tag{6.8}$$

where, $\mu$ represents the mean operator, and $\mathbf{X}_c$ represents all the training samples belonging to class $c$.

As discussed earlier in this section, we postulate that encoding the difference between the feature of a sample and the mean sample of the same class can help in encoding class-specific features. In other words, the feature of a particular class is brought similar/closer to the *mean* feature of that class. To encode this information, Eqs. 6.7 and 6.8 are utilized to form the following optimization constraint:

$$\|f_{x_c} - m_c\|_2^2 \tag{6.9}$$

The above equation is incorporated into an autoencoder to create Class Specific Mean Autoencoder as follows:

$$\underset{\mathbf{W}_e, \mathbf{W}_d}{\operatorname{argmin}} \|x_c - \mathbf{W}_d \phi(\mathbf{W}_e x_c)\|_2^2 + \lambda \|f_{x_c} - m_c\|_2^2 \tag{6.10}$$

where, $\lambda$ is the regularization constant. The proposed Class Specific Mean Autoencoder learns the weight parameters such that the features of a particular class are *grouped* together. Expanding Equation 6.10, we obtain:

$$\underset{\mathbf{W}_e, \mathbf{W}_d}{\operatorname{argmin}} \|x_c - \mathbf{W}_d \phi(\mathbf{W}_e x_c)\|_2^2 + \lambda \|\phi(\mathbf{W}_e x_c) - \mu(\phi(\mathbf{W}_e \mathbf{X}_c))\|_2^2 \tag{6.11}$$

The updated loss function of Equation 6.11 ensures that the learned feature for a sample is close

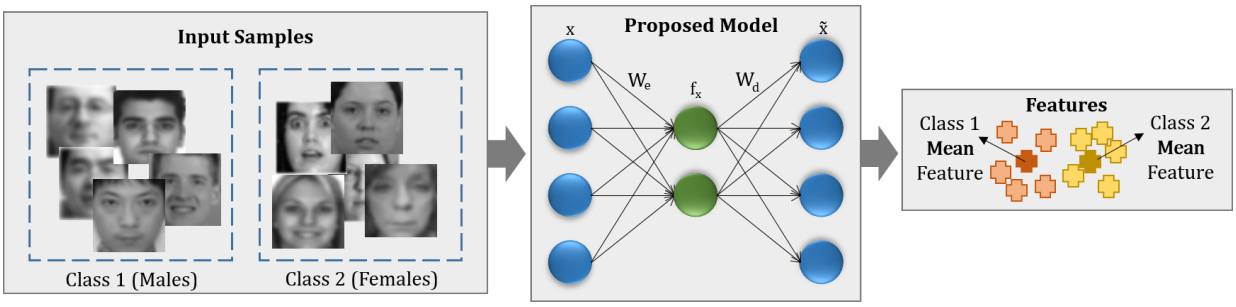



Figure 6-3: Proposed Class Specific Mean Autoencoder. $x$ and $\tilde{x}$ represent the input and the reconstructed samples respectively, $\mathbf{W}_e$ and $\mathbf{W}_d$ denote the encoding and decoding weights, and $f_x$ corresponds to the learned feature vector.

to the mean representation of its class, while being representative of the input sample as well. The second term is added for supervised regularization and can be viewed as:

$$E = \|f_{x_c} - t\|_2^2 \tag{6.12}$$

for a given expected target $t$ and obtained output $f_{x_c}$. The above equation draws a direct parallel with Equation 6.1, where the expected target is $x$, and the obtained output is $(\mathbf{W}_d\phi(\mathbf{W}_e x))$. Similar to the update rule for Equation 6.1, the update rule for the above regularization term for $j^{th}$ expected $(t_j)$ and obtained $(o_j)$ output, with respect to weight $w_{e_{i,j}}$, can be written as:

$$\frac{\partial \mathbf{E}}{\partial \mathbf{W}_{e_{i,j}}} = \frac{1}{2} * (o_j - t_j) * \frac{\partial o_j}{\partial \mathbf{W}_{e_{i,j}}} \tag{6.13}$$

Similar to the gradient descent backpropogation applied to Equation 6.1, the Class Specific Mean Autoencoder is solved iteratively via the above update rule till convergence.

For a $k$ layered Class Specific Mean Autoencoder, having encoding weights as $\mathbf{W}_e^1$, $\mathbf{W}_e^2$, ..., $\mathbf{W}_e^k$, and decoding weights as $\mathbf{W}_d^1$, $\mathbf{W}_d^2$, ..., $\mathbf{W}_d^k$, the loss function of Equation 6.10 can be modified as:

$$\underset{\substack{\mathbf{W}_e^1,\ldots,\mathbf{W}_e^k,\\ \mathbf{W}_d^1,\ldots,\mathbf{W}_d^k}}{\operatorname{argmin}} \|x_c - b \circ a(x_c)\|_2^2 + \sum_{i=1}^{i=k} \lambda_i \left\|f_{x_c}^i - m_c^i\right\|_2^2 \tag{6.14}$$

where, $a(x) = \phi(\mathbf{W}_e^k(\phi(\mathbf{W}_e^{k-1} \ldots (\phi(\mathbf{W}_e^1 x)))))$ is the encoding function, and $b(x) = \mathbf{W}_d^1(\mathbf{W}_d^2 \ldots (\mathbf{W}_d^k x))$

corresponds to the decoding function, and $f_{x_c}^k$ and $m_c^k$ are defined as:

$$f_{x_c}^k = \phi(\mathbf{W}_e^k(\phi(\mathbf{W}_e^{k-1} \dots (\phi(\mathbf{W}_e^1 x_c)))))) \tag{6.15}$$

$$m_c^k = \mu(\phi(\mathbf{W}_e^k(\phi(\mathbf{W}_e^{k-1} \dots (\phi(\mathbf{W}_e^1 \mathbf{X}_c)))))) \tag{6.16}$$

Owing to the large number of parameters involved, the optimization of the above model is performed via the greedy layer by layer approach [10]. At the time of testing, the learned encoding weights ($\mathbf{W}_e^1, \mathbf{W}_e^2, \dots \mathbf{W}_e^k$) are used to calculate the feature vector for a given sample, which is then provided as input to a classifier. Figure 6-3 presents a pictorial representation of the proposed algorithm, for a two class problem.

**Input :** Training images of male ($\mathbf{X}_{male}$) and female ($\mathbf{X}_{female}$) classes, iter = 0, maxIter.
**Output:** Encoding and decoding weights: $\mathbf{W}_e$, $\mathbf{W}_d$.
Initialize $\mathbf{W}_e$ and $\mathbf{W}_d$;
**while** *iter* $<$ *maxIter* **do**
  Compute mean male feature ($m_{male}^{iter}$) using Equation 6.8 ;
  Compute mean female feature ($m_{female}^{iter}$) using Equation 6.8 ;
  **foreach** $x_{female} \in \mathbf{X}_{female}$ **do**
    Minimize Equation 6.10 using $x_{female}$ and $m_{female}^{iter}$;
  **end**
  **foreach** $x_{male} \in \mathbf{X}_{male}$ **do**
    Minimize Equation 6.10 using $x_{male}$ and $m_{male}^{iter}$;
  **end**
  iter++;
**end**

**Algorithm 1:** Training Single Layer Class Specific Mean Autoencoder for Gender Prediction

### 6.2.3 Proposed Class Representative Autoencoder

The Class Specific Mean Autoencoder discussed above learns representations while modeling the intra-class variations only. However, the problem of gender classification is marred by the combined problem of high intra-class variations and low inter-class variations. To this effect, Class Representative Autoencoder is proposed which builds over the traditional unsupervised autoencoder, modeling both inter-class and intra-class variations at the time of feature learning. Gender

prediction is performed by the proposed model, termed as *AutoGen*. The optimization function incorporates class-specific terms for discriminative feature learning, in order to learn class-specific representations. This is done by minimizing the distance between a sample's representation and the representative feature of the sample's class, while maximizing the distance from the representative feature of all other classes. The class representative feature is computed as the mean feature vector of all the samples of a given class. The additional terms aid in incorporating inter-class and intra-class variations during the feature learning process such that the learned features are discriminative in nature. In a $n$ class problem, for a sample $x_c$ belonging to class $c$, the proposed model can be written as:

$$\underset{\theta}{\operatorname{argmin}} \|x_c - b \circ a(x_c)\|_F^2 + \lambda_s \|r_{x_c} - mean_c\|_F^2 - \sum_{i=1}^{n} \lambda_i \|r_{x_c} - mean_i\|_F^2 ; \quad \forall i \neq c \quad (6.17)$$

where, $\theta = \{\mathbf{W}_d, \mathbf{W}_e\}$, $r_{x_c}$ refers to the hidden representation of sample $x_c$, and $\lambda_s$ and $\lambda_i$ refer to the regularization constants for the additional terms. $mean_i$ refers to the mean feature (hidden) representation of all samples belonging to class $i$. For a single layer AutoGen, it can be computed as:

$$mean_i = \mu(\phi(\mathbf{W}_e \mathbf{X}_i)) \quad (6.18)$$

where, $\mathbf{W}_e$ are the encoding weights, $\mathbf{X}_i$ corresponds to all the training samples belonging to the $i^{th}$ class, and $\mu$ refers to the mean operator. Expanding Equation 6.17:

$$\begin{aligned} \underset{\theta}{\operatorname{argmin}} \|x_c - \mathbf{W}_d \phi(\mathbf{W}_e x_c)\|_F^2 + \lambda_s \|\phi(\mathbf{W}_e x_c) - \mu(\phi(\mathbf{W}_e \mathbf{X}_c))\|_F^2 \\ - \sum_{i=1}^{n} \lambda_i \|\phi(\mathbf{W}_e x_c) - \mu(\phi(\mathbf{W}_e \mathbf{X}_i))\|_F^2 \, \forall i \neq c \end{aligned} \quad (6.19)$$

Thus, the proposed formulation (Equation 6.19) consists of three terms: a term for learning the sample's feature, second for incorporating class similarity, and third for incorporating dissimilarity with other classes. As is the case with the traditional autoencoder, the first term aims to reduce the reconstruction error. The second term is responsible for bringing the learned representation ($r_{x_c}$) of a sample $x$ belonging to the class $c$ closer to $mean_c$ (representative feature of class $s$). Since the entire loss function is minimized, this *intra-class* term is also minimized, thus resulting

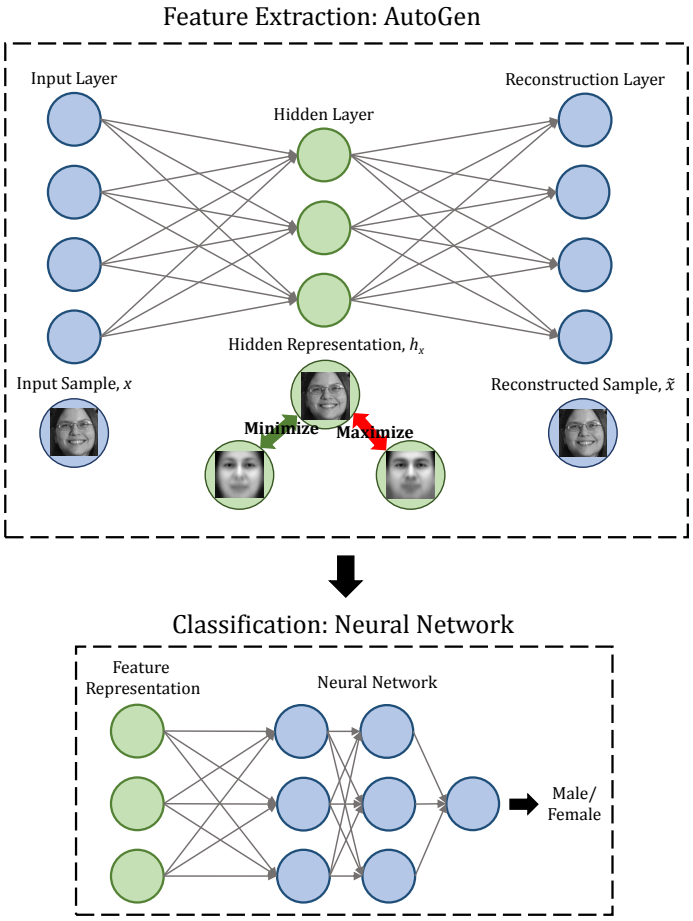



Figure 6-4: Pipeline for performing gender recognition on low resolution face images. The proposed AutoGen model is used for feature extraction, followed by a neural network for classification. AutoGen aims to learn discriminative features by incorporating inter-class and intra-class variations during feature learning. Figure has been taken from the published manuscript [137].

in learning feature vectors closer to the mean feature representation of that class. The third term in Equation 6.17 is responsible for maximizing the distance between the learned feature vector from the representative features of all other classes. This *inter-class* term attempts to force the model to learn a feature representation different from the representative (mean) feature of all other classes, thereby incorporating discriminability in the feature learning process.

In order to train the above model, back-propagation is performed using the gradient descent approach [48]. The model is trained for $k$ iterations, such that at the $j^{th}$ iteration, the parameters learned in the previous $(j-1)^{th}$ iteration are updated using the derivative of the loss function. The

weight update rule can be written as:

$$\mathbf{W^j} = \mathbf{W^{j-1}} - \eta \frac{\partial \mathbf{E}}{\partial \mathbf{W}} \tag{6.20}$$

where, $\mathbf{E}$ corresponds to the loss function given in Equation 6.19. Since all three terms in the formulation are differentiable (when using a differentiable activation function), the rule for updating the weights at each iteration can be obtained by calculating the derivative of each term. In order to extend the algorithm for deeper layers, Class Representative Autoencoder can be trained in a greedy layer-by-layer manner [10].

**AutoGen: Class Representative Autoencoder for Low Resolution Gender Recognition**

The proposed formulation aims to learn discriminative features in order to enhance classification of a given face sample into one of the two classes - male or female. At the time of training, mean representations for both the classes are calculated and incorporated by AutoGen for feature learning. The loss function, for a male input sample can be written as follows:

$$\underset{\theta}{\operatorname{argmin}} \|x_m - g \circ f(x_m)\|_F^2 + \lambda_m \|r_{x_m} - mean_m\|_F^2 - \lambda_f \|r_{x_m} - mean_f\|_F^2 \tag{6.21}$$

where, $r_{x_m}$ is the learned representation for the male input sample $x_m$, and $mean_m$ is the mean of features of all the male samples. $mean_f$ corresponds to the mean of feature vector of all the female samples. Similarly, the loss function for a female input sample is as follows:

$$\underset{\theta}{\operatorname{argmin}} \|x_f - g \circ f(x_f)\|_F^2 + \lambda_f \left\|r_{x_f} - mean_f\right\|_F^2 - \lambda_m \left\|r_{x_f} - mean_m\right\|_F^2 \tag{6.22}$$

where, $r_{x_f}$ is the learned representation for the female input sample $x_f$. The additional terms in the proposed model contribute to the supervised regularization during the feature learning process. The proposed model is used for feature extraction, which is followed by a neural network for classification. Figure 6-4 illustrates a pictorial representation of the proposed algorithm. At the time of testing, the learned encoding weights of AutoGen ($\mathbf{W}_e$) are used to obtain the feature representation of the given test face image. The feature is then provided as input to the trained

neural network, which finally predicts whether the input face image belongs to a *male* class or a *female* class.

## 6.3 Datasets and Experimental Protocol

This research aims to address gender classification in low resolution face images. In order to evaluate the performance of the proposed algorithm, the following datasets are used that contain images at different resolutions.

### 6.3.1 Datasets Used

Experiments are performed on the CMU Multi-PIE [44] and SCface [42] datasets. Details of each dataset are as follows:

**CMU Multi-PIE Dataset [44]:** The dataset contains images of 337 subjects, captured in indoor settings having variations with respect to pose, illumination, and expression. A frontal only subset of the dataset containing 50,248 images is used in the experiments. The dataset is divided into training and testing partitions, such that the train set consists of 18,420 images and the remaining 31,828 images make up the test set. Equal samples from both classes is ensured in the training partition.

**SCface Dataset [42]:** This dataset consists of visible and NIR images of 130 subjects captured over three distances. For each distance, one subject has five visible images and two NIR images captured in uncontrolled indoor environment. Since the aim is to perform low-resolution gender recognition, images pertaining to only the farthest distance have been used in the experiments. For visible spectrum experiments, 100 images are used for training, and 550 are used for testing. Exclusivity of subjects across the training and testing partition is maintained.

### 6.3.2 Experimental Protocol

As part of pre-processing, face detection is performed on all datasets using Viola Jones face detector [163], which is followed by geometric normalization of the face images. The detected faces are then resized to specific resolutions for different experiments. For all the datasets, the training

Table 6.2: Classification accuracies (%) for gender classification on $24 \times 24$ and $16 \times 16$ face images from the CMU Multi-PIE dataset.

| **Algorithm** | $\mathbf{24 \times 24}$ | $\mathbf{16 \times 16}$ |
|---|---|---|
| Autoencoder (AE) | 88.80 | 88.80 |
| Denoising AE | 88.21 | 87.82 |
| Deep Belief Network | 79.99 | 72.36 |
| Discriminative RBM | 72.84 | 70.41 |
| COTS: Face++ | 73.93 | 0.00 |
| COTS: Luxand | 74.14 | 0.00 |
| **Class Specific Mean AE** | **89.86** | **89.54** |
| **Class Representative AE (AutoGen)** | **90.10** | **89.57** |

partition is used to train the feature learning models and classifier, while testing is performed on the test set. For the CMU Multi-PIE dataset, experiments are performed for two resolutions: $24 \times 24$ and $16 \times 16$, while since only the farthest distance is chosen for the SCface dataset, experiments are performed at $24 \times 24$ resolution. The proposed autoencoder is trained on the training set of the CMU Multi-PIE dataset, which is fine-tuned for the SCface dataset.

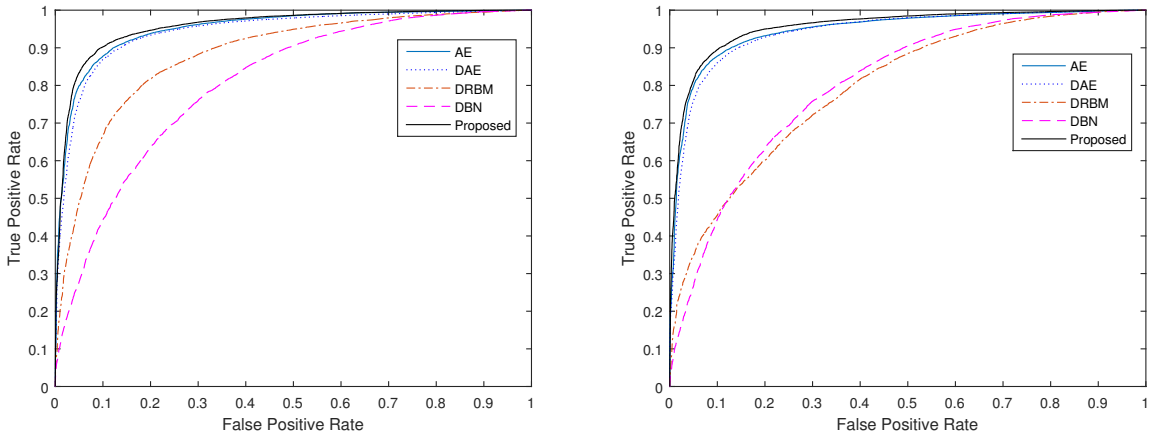
## 6.4 Results and Analysis

In order to compare the performance of the proposed model with existing algorithms, comparison has been drawn with Autoencoders (AE) [49], Denoising Autoencoders (DAE) [161], Deep Belief Networks (DBN) [52], Discriminative Restricted Boltzmann Machine (DRBM) [75], and two Commercial-Off-The-Shelf (COTS) systems, Luxand [1] and CNN-based Face++ [197]. All models are trained with a fixed architecture of $[k, k]$, where $k$ is the size of the input image. Once the feature learning process is over, features are extracted using the model and a Neural Network is trained for the 2-class classification problem of male versus female. A two layer neural network of dimensionality $[l, \frac{l}{4}, \frac{l}{16}]$ is trained, where $l$ is the length of the feature vector. $sigmoid$ activation function was used on all hidden layers. Due to the large class-imbalance of test samples, mean class-wise accuracy is reported for all the experiments.

**Classification Performance on the CMU Multi-PIE Dataset:** Table 6.2 presents the classification performance of the proposed models for $24 \times 24$ and $16 \times 16$ face images. For both the reso-

Table 6.3: Class specific classification accuracies (%) obtained with AutoGen for gender classification on the CMU Multi-PIE dataset.

| **Male** | **Female** |
|---|---|
| $24 \times 24$ | |
| 87.47 | 92.73 |
| $16 \times 16$ | |
| 86.50 | 92.64 |




(a) $24 \times 24$ (b) $16 \times 16$

Figure 6-5: Receiver Operating Characteristic (ROC) curves for $24 \times 24$ and $16 \times 16$ resolution face images from the CMU Multi-PIE dataset. Here, 'Proposed' refers to the AutoGen model.

lutions, the proposed autoencoder based models outperform existing algorithms and commercial-off-the-shelf systems by reporting higher classification accuracies. In case of $24 \times 24$, AutoGen gives a classification accuracy of 90.10%, which is at least 7% higher than COTS (Face++). For $16 \times 16$ resolution, a classification accuracy of 89.57% is achieved, which depicts improvement over other comparative algorithms. It is important to note that the commercial matchers, Face++ and Luxand fail to process any image of this resolution, thereby resulting in 0% classification performance. Figure 6-5 presents the Receiver Operative Characteristic (ROC) curves obtained for the experiments.

Table 6.3 also presents the class-wise classification accuracy obtained using AutoGen for the above experiments. It can be observed that for face images of resolution $24 \times 24$, female classification accuracy is significantly higher than male classification accuracy for both the resolutions.

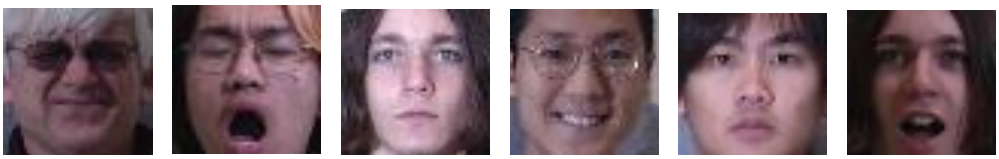


Figure 6-6: Sample male images misclassified as females by AutoGen. Most of the misclassifications are categorized by unusual hairstyle, expression, or accessories such as sunglasses.

Table 6.4: Classification performance (%) for gender classification on $24 \times 24$ face images, for the SCface dataset.

| **Algorithm** | **Accuracy (%)** |
|---|---|
| Autoencoder (AE) | 84.29 |
| Denoising AE | 81.35 |
| Deep Belief Network | 70.19 |
| COTS: Face++ | 2.72 |
| COTS: Luxand | 65.54 |
| **Class Specific Mean AE** | **87.82** |
| **Class Representative AE (AutoGen)** | **88.53** |

Upon analyzing the misclassified samples of males as females (Figure 6-6), it can be observed that unusual hairstyle, accessories, and sunglasses act as challenging artifacts for gender recognition. The performance of commercial-off-the-shelf systems, especially for $16 \times 16$ face images further reinstates the need for robust algorithms for gender recognition. It can also be observed that the accuracies obtained by AutoGen on $16 \times 16$ face images and $24 \times 24$ vary by less than 2%, for each dataset. This suggests that even with low resolution face images, the model is able to learn sufficient discriminative features.

**Data Captured in in Real-World Scenarios: SCface Dataset:** SCface dataset [42] has been captured in real world conditions containing 130 subjects. Since the number of training samples for the SCface dataset are very less, a fine-tuning approach was applied for the models. The trained feature extractor for CMU Multi-PIE dataset was fine-tuned with the training set of the SCface dataset. Table 6.6 gives the classification accuracy obtained on the dataset. It can be observed that the proposed models outperform existing algorithms by at least 4% by obtaining a classification accuracy of 88.53% (AutoGen). The improved performance of AutoGen with a small training set further motivates the usage of the proposed algorithm for the given task.

**Effect of Number of Layers:** To study the effect of deeper architectures for the purpose of gender

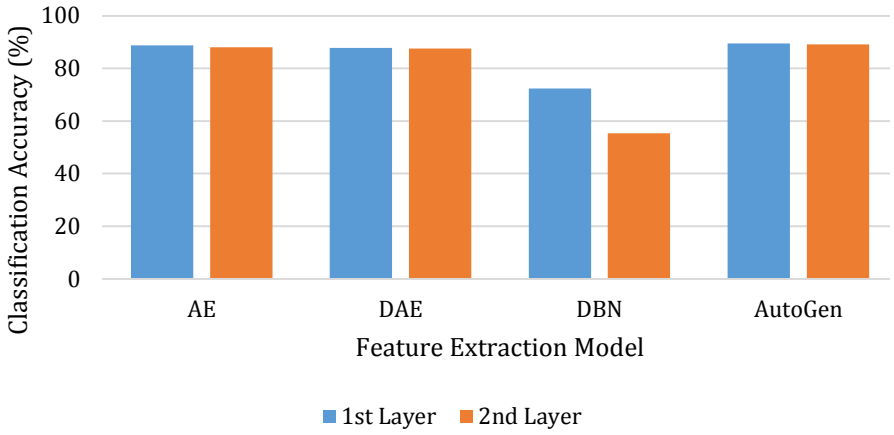



Figure 6-7: Comparison of classification accuracies for two layer feature extraction models, for face images having $16 \times 16$ resolution.

classification, all feature extraction models were trained for two layers. Greedy layer-wise training algorithm [10] was applied for $16 \times 16$ face images. Figure 6-7 presents the classification accuracies obtained after learning two layers of feature extraction models, along with the first layer classification results. It can be observed that the performance of all models reduce upon going deeper, or upon learning higher level of abstractions. While the accuracy of the proposed model reduces by less than 1.5%, the classification performance of DBN reduces by at most 17%. This demonstrates that while AutoGen can be used for learning deeper feature representations, however, for gender recognition in low resolution images, a single layer model performs best.

**Visualizations:** Figure 6-8 depicts some sample reconstructed images from CMU Multi-PIE dataset, obtained using AutoGen. It can be seen that the reconstructions of the mean male and female feature representations appear visually similar to the mean images. Along with that, the reconstructions of different samples demonstrate that the reconstructed face images are representative of the class to which the sample belongs. It can also be seen that the reconstructed image of a sample is closer to the mean reconstruction of the sample's class, as opposed to the other class. These visualizations along with the experimental results motivate the use of AutoGen for the task of gender classification on low resolution face images.

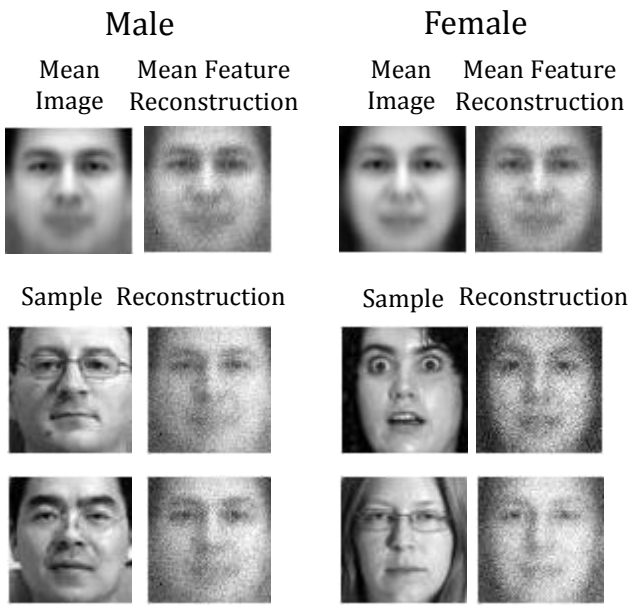



Figure 6-8: Sample reconstructed images from CMU Multi-PIE dataset using AutoGen.

## 6.5 Additional Experiments on Attribute Prediction

Beyond gender recognition from low resolution face images, the proposed models have also been evaluated on two additional case studies: (i) gender prediction from low resolution Near Infrared (NIR) images and (ii) adulthood prediction from face images. Details regarding each case study are given in the following subsections.

### 6.5.1 Gender Prediction from Low Resolution NIR Images

Extending upon the experiments discussed in the previous Section, the proposed models have also been evaluated from gender prediction from NIR images. The implementation details, comparative algorithms, and classification pipeline remain consistent as before. Experiments have been performed on two datasets:

- **PolyU-NIR Face Dataset [186]:** The PolyU NIR Face Dataset consists of 34,000 NIR spectrum face images of 335 subjects, having pose, illumination and distance variations. Out of these, 9,960 images are used for training (4,980 per class) and the remaining images form the test set.

- **SCface Dataset [42]:** This dataset consists of visible and NIR images of 130 subjects cap-

Table 6.5: Classification accuracies (%) for gender classification on $16 \times 16$ and $24 \times 24$ face images from the PolyU NIR dataset.

| **Algorithm** | $\mathbf{24 \times 24}$ | $\mathbf{16 \times 16}$ |
|---|---|---|
| Autoencoder (AE) | 63.46 | 61.28 |
| Denoising AE | 62.95 | 63.91 |
| Deep Belief Network | 65.10 | 49.00 |
| Discriminative RBM | 50.01 | 52.93 |
| COTS: Face++ | 52.46 | 0.00 |
| COTS: Luxand | 35.10 | 0.00 |
| **Class Representative AE (AutoGen)** | **71.32** | **69.86** |

tured over three distances. For each distance, one subject has five visible images and two NIR images captured in uncontrolled indoor environment. Images pertaining to only the farthest distance have been used in the experiments. For NIR spectrum experiments, 52 images are used for training and the remaining form the test set. Exclusivity of subjects across the training and testing partition is maintained.

**Experimental Protocol:** As part of pre-processing, face detection is performed on all datasets using Viola Jones face detector [163], which is followed by geometric normalization of the face images. The detected faces are then resized to specific resolutions for different experiments. For all the datasets, the training partition is used to train the feature learning models and classifier, while testing is performed on the test set. For the PolyU-NIR dataset, experiments are performed for two resolution: $24 \times 24$ and $16 \times 16$, while for the the SCface dataset, experiments are performed at $24 \times 24$ resolution only. For the SCface dataset, the proposed autoencoder is initially pre-trained on the PolyU-NIR dataset, followed by fine-tuning on the SCface dataset. An additional experiment is also performed for spectrum-invariant gender classification. In this case, the performance of a single model is analyzed for both NIR and visible images. The training sets of CMU Multi-PIE (discussed in the previous Section) and PolyU-NIR datasets are combined to create a single multi-spectrum training set. Gender classification performance is reported with the proposed AutoGen model on the testing partitions of the two datasets.

**Results and Analysis for Gender Prediction from Low Resolution NIR Images:** Table 6.5 and Table 6.6 present the classification accuracies obtained on the PolyU-NIR dataset and the SCface dataset, respectively. For both the resolutions and across datasets, the proposed AutoGen model

Table 6.6: Classification accuracies (%) for gender classification on $24 \times 24$ NIR face images, for SCface dataset.

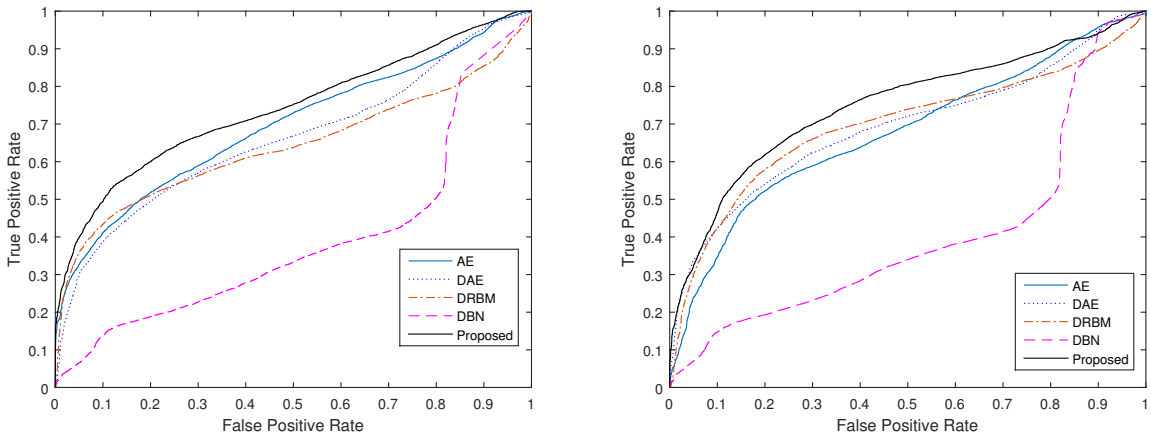
| **Algorithm** | **Accuracy (%)** |
|---|---|
| Autoencoder (AE) | 91.83 |
| Denoising AE | 88.61 |
| Deep Belief Network | 50.00 |
| COTS: Face++ | 0.00 |
| COTS: Luxand | 15.26 |
| **Class Representative AE (AutoGen)** | **95.79** |

Table 6.7: Class specific classification accuracies (%) obtained with **AutoGen** for gender classification on the PolyU NIR dataset.

| **Male** | **Female** |
|---|---|
| $24 \times 24$ | |
| 66.47 | 76.17 |
| $16 \times 16$ | |
| 70.47 | 69.24 |

outperforms the comparative algorithms and commercial-off-the-shelf systems. For the PolyU-NIR face dataset, AutoGen presents a classification accuracy of 71.32% and 69.86% on the two resolutions, which gives an improvement of at least 5% over existing algorithms, and at least 20% over COTS. Figure 6-9 presents the Receiver Operative Characteristic (ROC) curves obtained for the experiments. Table 6.7 also presents the class-wise classification accuracy obtained using AutoGen on the PolyU-NIR dataset. It can be observed that for face images of resolution $24 \times 24$, female classification accuracy is significantly higher than male classification accuracy.

**Training AutoGen for Spectrum Invariant Gender Classification:** In an attempt to model the gender variations across spectra in a single architecture, AutoGen was trained for face images of both spectrum, having a resolution of $16 \times 16$ pixels. A single model of AutoGen was trained using images of NIR and visible spectrum from the training sets of CMU Multi-PIE and PolyU-NIR. A classification accuracy of 88.47% and 69.66% is obtained for CMU Multi-PIE and PolyU-NIR face datasets respectively. In case of visible spectrum, a drop of around 1.1% is observed in the recognition performance (88.47% as opposed to 89.57%), and a drop of less than a percent is observed in case of NIR gender classification. This experiment suggests that while a combined




(a) $24 \times 24$ (b) $16 \times 16$

Figure 6-9: Receiver Operating Characteristic (ROC) curves for $24 \times 24$ and $16 \times 16$ resolution face images for the PolyU NIR dataset. Here, 'Proposed' refers to the AutoGen model.

model does not display any improvement in the classification performances, however, the drop in accuracies is also very less. This motivates formulation of spectrum invariant models for the given task.

### 6.5.2 Adulthood Prediction

Human aging is a complex process and brings with it behavioral and physiological changes. One associates maturity and mental growth with the behavioral changes that occur with time. Age is also often used as a means of access control, physically as well as virtually, to keep younger minds away from activities and content they are not deemed ready for. A threshold age, known as the *age of majority*, is defined by most states to universalize the concept of an individual being physically and mentally ready to assume control for their actions and decisions. However, there are different age limits prescribed by individual state and federal governments for different activities. For instance, in the USA, the legal age for smoking is 18 years while age limit for voting and drinking is 21 years.

The physiological and behavioral effects of aging vary for every individual and are a function of several parameters such as health, living style, environmental conditions, and gender. Therefore, it is challenging to accurately estimate the age of a person. As can be seen from Figure 6-10, it becomes difficult to predict the age of individuals just immediately below, or above the age

(a) Below the age of majority

(b) At the age of majority

Figure 6-10: Sample images from the FG-NET Aging dataset [111]. (a) shows images of individuals below the age of majority, and (b) shows sample individuals at the age of majority. These examples illustrate the challenging nature of *adulthood classification*.

of majority (e.g., 18 years). Among the currently available non-intrusive biometrics, the face *changes* significantly with age and is therefore a preferred modality for estimating the age of a person. As of now, at different checkpoints, an officer or a designated person in-charge estimates the age of a person by visually observing an individual. In cases where visual inspection becomes difficult, she/he asks for an identification (ID) card as the proof of age. However, research has shown that both these measures used for age estimation and verification are prone to errors [32, 93]. With easy availability of tampered or fake ID cards, several youngsters use these cards to mis-represent their identity. Recently, in a survey at Harvard University [172], it was found that $18\%$ underage students obtained alcohol using fake ID cards. Inspired by these observations and motivated by the increasing use of technology and automation in our day-to-day life, this research aims to automate the process of classifying an individual as an adult or not. For a given face image, the proposed algorithm aims to predict if the person has attained the age of majority or not. Such a system could be deployed at multiple places having age based restricted access; for instance, voting centers, driving license centers, and traffic check posts to scan for minors, restaurants and bars to prevent under-age alcohol consumption, cinema halls to enable restricted access, or around tobacco selling vending machines. Apart from the above mentioned applications, such systems can also be deployed virtually, where access is granted based on the age; for example, in online poker

rooms. To this effect, the proposed autoencoder based formulations have also been evaluated for the challenging task of adulthood prediction from face images.

**Datasets and Experimental Details:** Given a face image, predicting whether the individual is an adult or not, can be modeled as a two class classification problem: individuals below the age of 18 years are referred as *minors*, while individuals of age equal to or greater than 18 years are referred as *adults*. Experiments are performed on two datasets: (i) Multi-Ethnicity dataset and (ii) MORPH Album-II dataset. Figure 6-11 shows some sample images from both the datasets. Details about each are given below:

- Multi-Ethnicity Dataset: Since the existing datasets containing images of minors and adults comprise very limited variations with respect to ethnicity, pose, and expression, along with very few samples below the age of 16, we propose the Multi-Ethnicity Dataset. The dataset contains variations across ethnicity, gender, resolution, illumination, as well as minute pose and expression. It consists of 13,133 face images obtained from: (i) proposed Multi-Resolution Face Dataset containing 4,019 face images, and (ii) Heterogeneous Dataset containing 8,112 face images [25], and (iii) FG-Net Aging Dataset containing 1,002 face images [111]. Due to the lack of datasets containing face images of both adults and minors, we have created the *Multi-Resolution Face Dataset (MRFD)* consisting of 4,019 face images of minors and adults of two resolutions with slight variations in pose, expression, and illumination. The Multi-Resolution Face Dataset consists of face images of 317 Indian subjects captured in outdoor, as well as indoor environment. The dataset consists of images of 307 minors (3,896 images) and 10 adults. Images have been captured from two smartphones (with 3.1 MP camera) with resulting face size of $360 \times 420$ pixels, and a high resolution hand-held Canon digital camera with resulting face images of dimension $560 \times 680$. Each subject has at least 12 near-frontal, well-illuminated images (at least 4 from each camera source). The dataset contains variations in age, ranging from toddlers to adults of around 50 years. The subjects were only asked to look at the camera, without any instructions for pose or expression, which resulted in images with varying head movement and expression. This is the first dataset containing such large number of minor images which would help in facilitating research on minor face images as well.

(a) Multi-Ethnicity Dataset

(b) MORPH Album-II Dataset

Figure 6-11: Sample images from the datasets used for experimental evaluation.

Table 6.8: Summarizing the dataset description and experimental protocol.

| **Dataset** | **Images** | **Number of Images of** | | **Number of Images in** | |
|---|---|---|---|---|---|
| | | **Minors** | **Adults** | **Train Set** | **Test Set** |
| MORPH Album-II Dataset | 55,132 | 3,330 | 51,802 | 4,662 | 50,470 |
| Multi-Ethnicity Dataset | 13,133 | 8,574 | 4,559 | 6,276 | 6,857 |

- MORPH Album-II Dataset: Craniofacial Longitudinal Morphological Face (MORPH) dataset [126] consists of two albums: Album-I contains scanned digital face images, while Album-II contains longitudinal digital face images captured over several years. A subset of Album-II containing 55,134 images of 13,000 subjects is made available for academic researchers, which has been used for experimental analysis in this research. The dataset contains images of subjects between the age range of 16 to 77 years, and also provides metadata for race, gender, date of birth, and date of acquisition.

**Experimental Protocol and Implementation Details:** Unseen training and testing partitions are created for both the datasets. For training, equal number of samples from both the classes are used, which is defined by the class with lesser number of samples. 70% of the samples corresponding to the minor class and equal number of images from adult class are used for training, while the remaining data is used for testing. For the MORPH dataset, this results in the training and testing sets of size 4,662 and 50,470 images respectively. Similarly, for the Multi-Ethnicity dataset, 6,276 images are selected for training with the constraint that equal number of samples are selected from

both the classes. The remaining face images constitute the test set. Details of data partitioning are documented in Table 6.8.

To showcase the efficacy of the proposed algorithm, comparison has been drawn with other deep learning based feature extractors; namely, Stacked Denoising Autoencoder (SDAE) [162], Deep Boltzmann Machine (DBM) [50], and Discriminative Restricted Boltzmann Machine (DRBM) [75]. Comparison has also been drawn with VGG-Face descriptor [113], which is one of the state-of-the-art deep learning based feature extractor. Features extracted from these models are provided as input to a Neural Network for classification. A CNN based Commercial-Off-The-Shelf (COTS) system, Face++ [197], has also been used to compare the performance of the proposed model. Since there does not exist any COTS for the task of adulthood prediction, Face++ is used to estimate the age of the given face image, which is then utilized to classify the input as an adult or a minor. In order to analyze the statistical significance of the results obtained by the proposed model, McNemar test [96] has been performed. Given the classification results obtained from two models, McNemar test predicts whether the performance of both the models is statistically different or not. For every comparison of the proposed Class Specific Mean Autoencoder with an existing architecture, a $p$-value is reported. A smaller $p$-value corresponds to a higher confidence level of statistical difference. In this research, all claims of statistical significance have been made at a confidence level of 95%.

For all the experiments, face detection is performed on all images using Viola Jones Face Detector [163], following which the images are geometrically normalized and resized to a fixed size. A Class Specific Mean Autoencoder of dimensions [$m$, $m$] is learned, where $m$ is the size of the image. Following this, a neural network of dimension [$\frac{m}{4}$, $\frac{m}{8}$] is trained for classification. $sigmoid$ activation function is used at the hidden layers. Models are trained for 100 epochs with a learning rate of 0.01. We have followed the best practices used for setting the parameters and architecture for deep learning [76]. For existing algorithms, in order to maintain consistency, a two layer architecture is utilized for the feature extractor and neural network.

**Experimental Results and Observations:** Owing to the large class imbalance in the test samples, mean class-wise accuracy has been reported for all the experiments. The formula used for

calculating the accuracy is as follows:

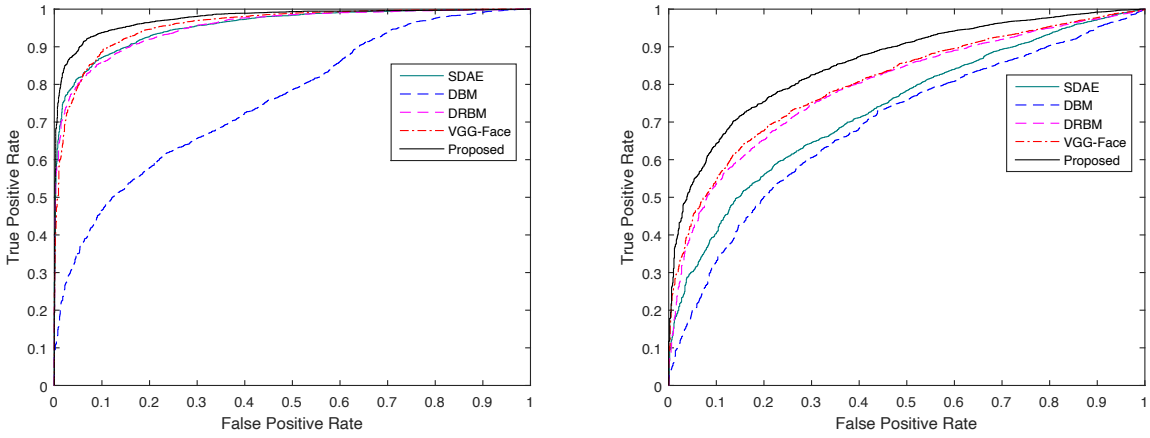
$$Accuracy = \frac{Accuracy_{Minor} + Accuracy_{Adult}}{2} \tag{6.23}$$

where, $Accuracy_{Minor}$ and $Accuracy_{Adult}$ correspond to the accuracies obtained for minor and adult classification, respectively by a particular model.

**Results on Multi-Ethnicity Dataset:** Table 6.9 presents the classification accuracies of the proposed model along with other existing architectures on the Multi-Ethnicity Dataset. Figure 6-12 also presents the Receiver Operative Characteristic (ROC) curves obtained for the experiments. It is observed that the proposed Class Specific Mean Autoencoder (2-layer) achieves a classification accuracy of **92.09%**, which is at least 2.5% better than existing algorithms. This is followed by VGG-Face with an accuracy of 89.45%, while Face++ (commercial off-the-shelf system) achieves a classification performance of 78.41%. The improvement of 5.29% in performance of the proposed model as compared to a Stacked Denoising Autoencoder can be attributed to the additional class representative terms added to the autoencoder formulation.

Table 6.9 also presents the $p$-values obtained upon performing the McNemar test to evaluate the statistical difference. Since all the values are below 0.05, we can claim with a confidence level of 95% that the performance of the proposed model is statistically different from all other existing models. In order to understand the effect of number of layers, experiments are also performed using a single layer Class Specific Mean Autoencoder. For a single layer, the proposed model yields an accuracy of 91.58%, which continues to show an improvement of at least 2%, compared to other models with the same architecture.

To analyze the class-specific classification accuracies, Table 6.11 presents the confusion matrix for the proposed Class Specific Mean Autoencoder on the Multi-Ethnicity dataset. The results indicate that the performance of the trained model is not biased towards any particular class by achieving a classification accuracy of 93.52% and 90.65% on the two classes of adults and minors, respectively. This is essential to ensure that while unauthorized access is not provided to minors, rightful adults are not restricted from it either. In order to cater to the application of age-specific authorized access control, it is essential to ensure that the percentage of people below the age of majority i.e., minor, obtaining unauthorized access should be minimal. To analyze the perfor-




(a) Multi-Ethnicity Dataset (b) MORPH Album-II Dataset

Figure 6-12: Receiver Operating Characteristic (ROC) curves obtained for categorizing whether a given face image is of an adult or not. Here, 'Proposed' refers to the Class Specific Mean Autoencoder. Figure has been taken from the published manuscript [141].

Figure 6-13: Sample images from the Multi-Ethnicity dataset, incorrectly classified by all algorithms. At the time of capture, all individuals were below the age of 18. It can be seen that while the actual age of the samples was below the age of majority, it is easy to mistake minors of 16-17 years as adults. External accessories such as scarves may also introduce mis-classification, resulting in unauthorized access control.

mance of all architectures for such an application, Figure 6-15 presents bar graphs summarizing the percentage of minors misclassified as adults. It can be seen that the proposed model achieves a misclassification percentage of 9.35%, as opposed to 22.97% by Face++. Figure 6-13 presents some sample images from the Multi-Ethnicity dataset misclassified as adults by *all* the algorithms. It can be observed from the sample images that these images were captured either near the age of majority of 16-17 years or have artifacts such as headbands/scarves, which further make the task of adulthood classification challenging. Certain samples of kids below the age of one year were also mis-classified, possibly due to the undeveloped features of newborns.

The major challenges associated with the problem of adult classification lie in the age bracket

Table 6.9: Classification Accuracy (%) on Multi-Ethnicity dataset. $p$-Value corresponds to the values obtained after performing McNemar test to compare the classification performance of an existing architecture with the proposed Class Specific Mean Autoencoder. The proposed model presents improved classification performance, while being statistically different from all other models at a confidence level of 95%.

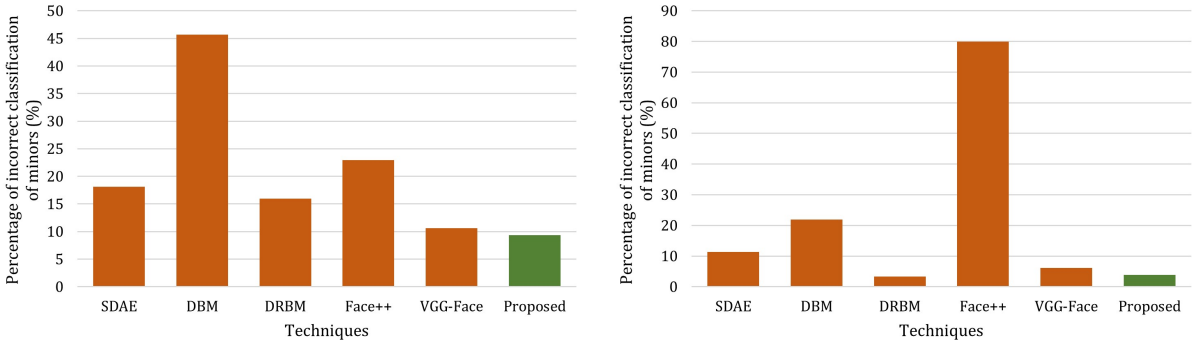
| **Method** | **Accuracy (%)** | $p$-**Value** | **Statistical Significance** |
|---|---|---|---|
| SDAE [162] | 86.80 | 0.003 | Significant |
| DBM [50] | 65.16 | $< 0.001$ | Significant |
| DRBM [75] | 87.03 | 0.001 | Significant |
| VGG-Face [113] | 89.45 | 0.004 | Significant |
| COTS: Face++ [197] | 78.41 | $< 0.001$ | Significant |
| **Class Specific Mean Autoencoder** | **92.09** | - | - |

Table 6.10: Classification Accuracy (%) on MORPH Album-II dataset. $p$-Value corresponds to the values obtained after performing McNemar test to compare the classification performance of an existing architecture with the proposed Class Specific Mean Autoencoder. The proposed model presents improved classification performance, while being statistically different from all other models at a confidence level of 95%.

| **Method** | **Accuracy (%)** | $p$-**Value** | **Statistical Significance** |
|---|---|---|---|
| SDAE [162] | 66.25 | 0.005 | Significant |
| DBM [50] | 65.30 | $< 0.001$ | Significant |
| DRBM [75] | 65.72 | $< 0.001$ | Significant |
| VGG-Face [113] | 70.44 | 0.010 | Significant |
| COTS: Face++ [197] | 57.23 | $< 0.001$ | Significant |
| **Class Specific Mean Autoencoder** | **73.13** | - | - |

Figure 6-14: Sample images from the Multi-Ethnicity dataset that are in the age bracket of 16-19 years and misclassified by the proposed Class Specific Mean Autoencoder. The first image belongs to an adult of age 19 years, while the remaining belong to entities below the age of majority.

of 16 to 19 years (16-17: minors, and 18-19: adults). On the Multi-Ethnicity dataset, the proposed algorithm achieves a classification accuracy of 64.58% on the above mentioned age range. VGG-Face, which performs the second best, reports an accuracy of 58.33%, which is at least 6% lower than the proposed algorithm. Figure 6-14 displays sample images from the specified age range and




(a) Multi-Ethnicity Dataset

(b) MORPH Album-II Dataset

Figure 6-15: Percentage of minors incorrectly classified as adults for both the datasets. A lower percentage would ensure fewer instances of unauthorized access. Here, 'Proposed' refers to the Class Specific Mean Autoencoder.

Table 6.11: Confusion matrix of Class Specific Mean Autoencoder on the Multi-Ethnicity database.

| | | **Predicted** | |
|---|---|---|---|
| **Actual** | | Adult | Not Adult |
| | Adult | 93.52% | 6.48% |
| | Not Adult | 9.35% | 90.65% |

are misclassified by the proposed algorithm. The images demonstrate the challenging nature of human aging which are dependent on intrinsic and extrinsic person-specific factors, such as health, environment, and climate.

**Results on MORPH Album-II Dataset:** The classification accuracies obtained by the proposed model and other existing architectures are tabulated in Table 6.10, and Figure 6-12 presents the Receiver Operating Characteristic (ROC) curves obtained for the experiments. It is observed that the proposed architecture achieves a classification accuracy of **73.13%**, which is at least 2.5% better than existing approaches, while Face++ (COTS) achieves an accuracy of 57.23%. Table 6.10 also presents the $p$-values obtained upon performing the McNemar statistical test on the proposed Class Specific Mean Autoencoder and other existing models. While the second best performance is achieved by VGG-Face features (70.44%), it is important to note that the improvement in accuracy achieved by the proposed model is statistically significant for a confidence level of 95%. Upon analyzing the gender-specific adulthood prediction results, it can be observed that the classifica-

(a) Minors (b) Adults

Figure 6-16: Sample images from the MORPH Album-II dataset correctly classified by the proposed algorithm, and not by other existing algorithms. (a) Images of individuals having age 16 (first two samples) or 17. (b) Images of just turned adults of 18 (first two) or 19 years of age.

Table 6.12: Classification accuracy (%) on perturbed face images for Multi-Ethnicity and MORPH Album-II datasets.

| **Perturbation** | **Multi-Ethnicity Dataset** | | **MORPH Dataset** | |
|---|---|---|---|---|
| | **Proposed** | **VGG-Face** | **Proposed** | **VGG-Face** |
| No Perturbation (Original) | 92.09 | 89.45 | 73.13 | 70.44 |
| Gaussian Blur (Sigma = 3) | 89.54 | 61.34 | 72.40 | 50.00 |
| Gaussian Noise ($\mu$ = 0, Std. dev. = 0.01) | 87.95 | 50.59 | 72.09 | 50.03 |
| Gaussian Noise($\mu$ = 0, Std. dev. = 0.001) | 91.66 | 60.57 | 72.98 | 62.08 |
| Holes (10 holes of 3 $\times$ 3) | 87.35 | 64.67 | 72.55 | 56.70 |

tion accuracy on female sample images is 62.89%, whereas the accuracy on male sample images is 75.09%. It is further observed that for females, the misclassification of adults as minors is much higher, as compared to males, thereby resulting in an overall lower classification performance.

From Figure 6-15, it can be observed that the proposed model achieves a *minor* misclassification percentage of only 3.9%, as opposed to nearly 80% by Face++ (COTS) on the MORPH Album-II dataset. The high misclassification rate of minors by Face++ reinstates the requirement for robust algorithms with the ability to process and analyze minor face images as well. It is important to note that the age of face images in MORPH Album-II dataset varies from 16 to 77 years. Thus, resulting in a further challenging dataset having multiple subjects *just below* the age of majority. The higher misclassification rate achieved can thus also be attributed to this challenging age range. It is also interesting to observe from Figure 6-15 that while DRBM achieves a lower misclassification of minors as adults (i.e., 3.30%), the overall classification accuracy of DRBM based approach is less than the proposed approach (Table 6.10). This further motivates the use of the proposed algorithm for ensuring rightful access to adults, while restricting minors.

Figure 6-16 presents sample images of the age group of 16-19 (16-17: minors, 18-19: adults)

years from the MORPH Album-II dataset which are correctly identified by the proposed algorithm and not by any other algorithm. Upon analyzing the mean images of both the classes, we observe a significant visual difference in the jaw area of minors and adults. As can be seen from Figure 6-16 as well, minors appear to have a tighter jaw line which is often not observed with adults. We believe that this variation has been encoded well by the proposed model among other features, resulting in superior performance.

**Performance on Perturbed Face Images:** It has often been observed in literature that the performance of deep models deteriorates in the presence of perturbations [40]. The proposed model has also been evaluated on perturbed face images in order to understand its vulnerabilities. This is performed by incorporating perturbations in the form of Gaussian blur, Gaussian noise, and holes in the original face images. Experiments are performed on the Multi-Ethnicity and the MORPH Album-II datasets with the protocols discussed earlier. The models are trained on unperturbed (original) images but the test images are perturbed. In this evaluation, no separate training is performed for the perturbed face images. Table 6.12 presents the classification accuracies obtained from the proposed Class Specific Mean Autoencoder, and the second best performing model, VGG-Face. It can be observed that with perturbed test images, the accuracy of the proposed model reduces by less than 5% and 1.04% for Multi-Ethnicity and MORPH Album-II datasets, respectively. On the other hand, VGG-Face demonstrates a drop of at least 24% and 8% on the two datasets, respectively. This experiment demonstrates the utility of the proposed model for performing classification under different kinds of perturbations.

## 6.6 Summary

Gender prediction from low resolution face images is a challenging and less explored area, with wide-spread application. This research proposes two novel supervised autoencoders for the given task: (i) Class Specific Mean Autoencoder and (ii) Class Representative Autoencoder (termed as *AutoGen*. Both the models utilize class information during training to learn discriminative features, useful for classification. While the Class Specific Mean Autoencoder focuses on learning representations closer to the class mean, AutoGen also promotes higher inter-class variation during learning. The performance of the two models has been demonstrated on low resolution face images

having $16 \times 16$ and $24 \times 24$ resolution. The proposed models demonstrate improved classification performance as compared to the existing techniques and commercial-off-the-shelf systems. The improved performance strengthens the need to incorporate class-specific discriminative information at the feature extraction stage, and motivates the use of the proposed models. Further, additional experiments have also been performed on other attribute classification tasks such as gender prediction in NIR spectrum and adulthood prediction on face images. Across different tasks, datasets, and resolutions, the proposed models demonstrate improved performance. This research has been published in Pattern Recognition Letters and the International Joint Conference on Neural Networks (IJCNN), 2017. Most of the images have been taken from the above published manuscripts [137, 141].

# Chapter 7

# Conclusion and Future Research

This dissertation focuses on two challenging and less explored covariates of unconstrained facial analysis: images captured under (i) low resolution and (ii) disguise variations. Both the covariates are commonly observed in facial images captured from a distance, and present unique set of challenges. As part of this dissertation, deep learning based facial analysis models have been proposed, which demonstrate improved performance on different standard benchmark datasets as compared to the state-of-the-art methods. The key contributions of this dissertation are:

- **Face Recognition Models for (Very) Low Resolution Face Images:** A novel *Dual Directed Capsule Network (DirectCapsNet)* has been proposed for very low resolution face recognition. DirectCapsNet utilizes a combination of capsule layers and convolutional layers for learning an accurate classifier for VLR images. Further, a novel *DeriveNet* model has also been proposed for VLR/LR face recognition, which utilizes a novel Derived-Margin softmax loss. The DeriveNet model incorporates a derived margin for modeling the inter-class variations for learning improved class boundaries. Both the models utilize auxiliary HR samples during training, and attempt to learn from their higher interpretive information content. The proposed models have been evaluated on several real-world datasets including the challenging UCCS dataset and the DroneSURF dataset, where improved performance has been obtained.

- **Disguised Faces in the Wild Datasets:** This dissertation also presents two novel datasets: (i) *Disguised Faces in the Wild 2018 (DFW2018)* and *Disguised Faces in the Wild 2019*

*(DFW2019)*. These are the first-of-a-kind dataset containing disguised images and impersonator images for a given subject. Further, the DFW2019 dataset also contains images corresponding to bridal accessories and variations due to plastic surgery. Both the datasets have been collected from the Internet, thus demonstrating wide variability with respect to the disguise accessories. Statistical details regarding the datasets, benchmark protocols, baseline results, and the additional analysis has been discussed in the dissertation. Further, both the datasets were released as part of the DFW Challenge and Workshop across two consecutive years. The results and analysis of the submissions of the challenges have also been discussed, demonstrating relatively lower performance at stricter False Acceptance Rates.

- **Disguise-Resilient Face Verification:** Finally, the *Disguise-Resilient (D-Res)* pipeline has been presented for face verification under disguise variations. A novel Disguise Encoder-Decoder Network has been presented which focuses on learning a disguise invariant facial representation. The DFW2018 and DFW2019 datasets have been used for evaluating the efficacy of the proposed model, where it achieves state-of-the-art performance. Further, to the best of our knowledge, this is the first research which discusses the relevance of face recognition in the intersection of low resolution and disguise variations. To this effect, baseline results have been reported on the standard benchmark protocols of the two datasets, along with the results of the proposed pipeline.

- **Gender Prediction Models for Low Resolution Face Images:** This dissertation presents one of the initial research on automating gender prediction from LR face images. Two novel autoencoder formulations have been presented: (i) *Class Specific Mean Autoencoder* and (ii) *Class Representative Autoencoder*. The original unsupervised autoencoder model is updated to incorporate the class information during model training for learning enhanced discriminative features. The efficacy of the proposed models has been demonstrated on different datasets and different resolutions.

While the research presented in this dissertation advances the literature and state-of-the-art in face recognition,there still exists substantial scope for improvement. Figure 7-1 presents a diagrammatic overview of the suggested future work by building upon the contributions of this dissertation. Specifically, some potential directions for future research are as follows:

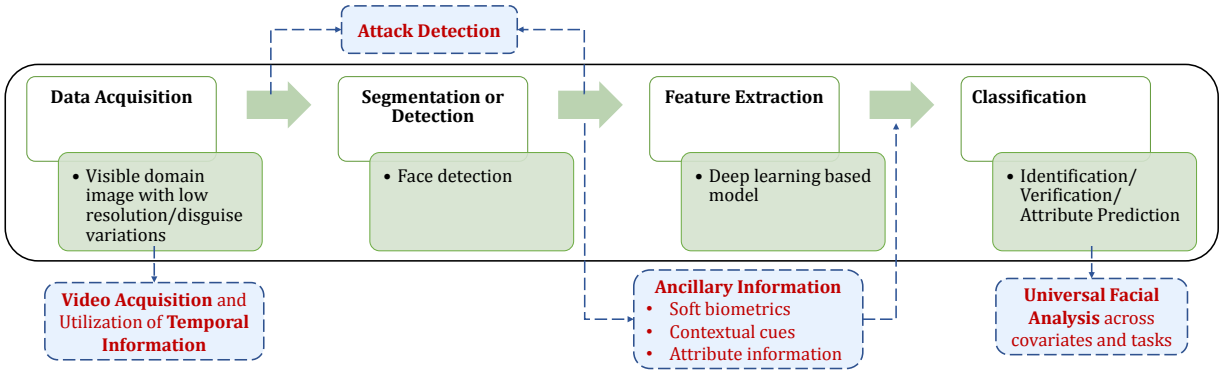



Figure 7-1: This dissertation presents algorithms for facial analysis under low resolution and disguise variations. Potential future research directions have been demonstrated in this figure including (i) utilizing temporal information for rich features via video acquisition, (ii) creating robust automated models invariant to adversarial attacks, (iii) incorporation of ancillary information for improved feature generation, and (iv) development of universal facial analysis models.

- **Developing a Universal Facial Analysis Framework:** Facial analysis algorithms deployed in real-world scenarios are expected to process unconstrained facial images captured under the effect of multiple covariates. As demonstrated in this dissertation, the performance of face recognition algorithms reduces significantly on low resolution images having disguise variations as compared to their high resolution counter-parts. The drop in performance thus suggests the need for developing *universal facial analysis* models or pipelines capable of handling the combination of different covariates at run-time. The domain of model ensemble could also be explored for achieving the desired goal, while maintaining the constraint of fast real-time processing. Further, the techniques presented in this dissertation for VLR/LR face recognition can also be extended to handle other variations observed in VLR/LR samples such as blurring, sensor noise, or compression artifacts, which render automated facial recognition further challenging. Beyond model developments, currently, the research community is also marred with the challenge of limited in-the-wild datasets capturing real-world variations.The DFW2018 and DFW2019 datasets, created as part of this dissertation, contain high resolution face images captured from the Internet with *disguised* and *impersonator* images. Since the datasets are collected from the Internet, the images mostly belong to well-known celebrities from across the World. While the dataset enabled further research, it can be extended to include facial images captured in real-time while maintaining the high variation of disguise accessories. Further, real-world datasets must also be created to cap-

ture multiple covariates together, as would be expected by a facial analysis model deployed in real-world settings. Dedicated research along the lines of algorithmic/model development and real-world datasets availability may ensure the usage of universal facial analysis frameworks in the near future. The following points further present potential solutions for strengthening facial analysis models.

- **Fusion of Ancillary Information for Accurate Face Recognition under Challenging Covariates:** This dissertation explored the broad area of attribute classification in low resolution face images, specifically, gender classification, where the proposed models achieved encouraging performance. Given the relatively lesser information content in low resolution samples, accurate attribute prediction can thus be utilized as *ancillary information* for developing more robust face recognition models, capable of high performance. Further, ancillary information such as gait, soft biometrics, and contextual cues can also aid in accurate face recognition in scenarios of obfuscated facial region as well. Several concepts of *biometric fusion* can be applied for the fusion of ancillary information into the face recognition pipeline, such as, feature-level fusion, score-level fusion, or decision-level fusion.

- **Developing (LR) Facial Analysis Models Robust to Attacks:** The development of face recognition algorithms for challenging scenarios such as low resolution and disguise variations also require them to be protected against unwanted attacks. While research has focused on evaluating and advancing the robustness of face recognition models against *adversarial attacks (perturbations)*, most of the research has catered towards models developed for high resolution face images only. Since noise is often an inherent characteristic of low resolution samples, a trained model might be invariant to slight pixel perturbations. However, since such models can have high applicability in critical real-world scenarios (e.g. recognition from a distance), it is imperative for upcoming research to focus on the challenging intersection of adversarial attacks and (very) low resolution face recognition in order to enhance the robustness of recognition models.

- **Video-based LR Facial Analysis:** One of the key applications of unconstrained facial analysis (especially from a distance) is for automated remote monitoring. In such scenarios, often a (CCTV) camera captures the live feed, which is processed at real-time. Here, the

live feed mostly corresponds to a video content as compared to images captured across varying timestamps. Therefore, a potential future research direction is to develop facial analysis algorithms for a video input, as opposed to an image based input. Here, the additional temporal information can help in developing robust algorithms capable of utilizing modalities beyond the face as well.

Finally, it is important to note that the work done as part of this dissertation and the proposed future work focus on utilizing facial analysis algorithms in an ethical manner for the betterment of the society by reducing the required manual effort in different facial analysis tasks. In scenarios where a face recognition system is marred with a large search space (millions of face identities), facial analysis algorithms focusing on attribute prediction can often be utilized as a pre-processing unit, thus helping in the reduction of the effective search space. Such combinations can be useful in scenarios of automated missing person identification, strengthening facial verification modules by adding an additional layer of attribute matching, or applying child-specific locks/restrictions on unmonitored devices such as smart TVs/iPads etc. It is our belief that facial analysis systems must be used in an ethical manner, in compliance with existing policies, and promote a better, safe, and an equitable life for all.

THIS PAGE INTENTIONALLY LEFT BLANK

# Bibliography


[1] Luxand. https://www.luxand.com. 130

[2] P. Afshar, A. Mohammadi, and K. N. Plataniotis. Brain tumor type classification via capsule networks. In *IEEE International Conference on Image Processing*, pages 3129–3133, 2018. 22

[3] O. A. Aghdam, B. Bozorgtabar, H. K. Ekenel, and J.-P. Thiran. Exploring factors for improving low resolution face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2363–2370, 2019. 38

[4] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 57

[5] G. Amato, F. Falchi, C. Gennaro, F. V. Massoli, and C. Vairo. Multi-resolution face recognition with drones. In *International Conference on Sensors, Signal and Image Processing*, 2020. XVI, 37, 52, 53

[6] Y. Andreu, J. LÃşpez-Centelles, R. A. Mollineda, and P. Garcia-Sevilla. Analysis of the effect of image resolution on automatic face gender classification. In *International Conference on Pattern Recognition*, pages 273–278, 2014. 118

[7] A. Bansal, R. Ranjan, C. D. Castillo, and R. Chellappa. Deep features for recognizing disguised faces in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–16, 2018. XVII, 68, 69, 72, 97

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recog-

nition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 1, 11

[9] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012. 106

[10] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*, pages 153–160. 2007. 125, 128, 133

[11] S. Bharadwaj, H. S. Bhatt, M. Vatsa, and R. Singh. Domain specific learning for newborn face recognition. *IEEE Transactions on Information Forensics and Security*, 11(7):1630–1641, 2016. 104

[12] H. S. Bhatt, R. Singh, M. Vatsa, and N. K. Ratha. Improving cross-resolution face matching using ensemble-based co-transfer learning. *IEEE Transactions on Image Processing*, 23(12):5654–5669, 2014. 20

[13] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer. Pose-robust recognition of low-resolution face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):3037–3049, 2013. XV, 31

[14] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. 2010. 45

[15] 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. 47, 59, 67, 82, 85

[16] C. Chen, A. Dantcheva, and A. Ross. An ensemble of patch-based subspaces for makeup-robust face recognition. *Information Fusion*, 32:80–92, 2016. 104

[17] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122, 2014. 119

[18] Z. Cheng, X. Zhu, and S. Gong. Surveillance face recognition challenge. *arXiv preprint arXiv:1804.09691*, 2018. 3, 54

[19] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 1

[20] 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. 41, 47, 48, 83, 84, 85, 110, 113

[21] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. RetinaFace: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 84, 85

[22] J. Deng and S. Zafeririou. Arcface for disguised face recognition. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 96, 97, 105, 107

[23] A. N. Desk. Top 10 countries and cities by number of CCTV cameras. 2019. 7

[24] T. I. Dhamecha, A. Nigam, R. Singh, and M. Vatsa. Disguise detection and face recognition in visible and thermal spectrums. In *International Conference on Biometrics*, 2013. 3, 11, 60

[25] T. I. Dhamecha, A. Sankaran, R. Singh, and M. Vatsa. Is gender classification across ethnicity feasible using discriminant functions? In *IEEE International Joint Conference on Biometrics*, pages 1–7, 2011. 139

[26] T. I. Dhamecha, R. Singh, M. Vatsa, and A. Kumar. Recognizing disguised faces: Human and machine evaluation. *PLOS ONE*, 9(7):1–16, 2014. XVII, 57, 62, 81, 97

[27] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *European Conference on Computer Vision*, pages 370–386, 2018. 39

[28] Y. Duan, J. Lu, J. Feng, and J. Zhou. Context-aware local binary feature learning for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1139–1153, 2018. 57

[29] Y. Duan, J. Lu, and J. Zhou. Uniformface: Learning deep equidistributed representation for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2019. 113

[30] S. S. Farfade, M. J. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks. In *International Conference on Multimedia Retrieval*, pages 643–650, 2015. 119

[31] A. Feingold. Gender differences in personality: a meta-analysis. *Psychological Bulletin*, 116(3):429, 1994. 117

[32] E. Ferguson and C. Wilkinson. Juvenile age estimation from facial images. *Science & Justice*, 57(1):58 – 62, 2017. 138

[33] J. Galbally, S. Marcel, and J. Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014. 57

[34] S. Gao, Y. Zhang, K. Jia, J. Lu, and Y. Zhang. Single sample face recognition via learning deep supervised autoencoders. *IEEE Transactions on Information Forensics and Security*, 10:2108–2118, 2015. 120, 121

[35] S. Ge, C. Li, S. Zhao, and D. Zeng. Occluded face recognition in the wild by identity-diversity inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3387–3397, 2020. 95

[36] S. Ge, K. Zhang, H. Liu, Y. Hua, S. Zhao, X. Jin, and H. Wen. Look one and more: Distilling hybrid order relational knowledge for cross-resolution image recognition. In *AAAI Conference on Artificial Intelligence*, pages 10845–10852, 2020. 51

[37] S. Ge, S. Zhao, C. Li, and J. Li. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing*, 28(4):2051–2062, 2019. 20, 27, 31, 38, 46, 51, 110

[38] S. Ge, S. Zhao, C. Li, Y. Zhang, and J. Li. Efficient low-resolution face recognition via bridge distillation. *IEEE Transactions on Image Processing*, 29:6898–6908, 2020. 51

[39] R. Giot and C. Rosenberger. A new soft biometric approach for keystroke dynamics based on gender recognition. *International Journal of Information Technology and Management*, 11(1-2):35–49, 2012. 117

[40] G. Goswami, N. K. Ratha, A. Agarwal, R. Singh, and M. Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *AAAI Conference on Artificial Intelligence*, 2018. 57, 147

[41] G. Goswami, M. Vatsa, and R. Singh. Face verification via learned representation on feature-rich video frames. *IEEE Transactions on Information Forensics and Security*, 12(7):1686–1698, 2017. 57

[42] M. Grgic, K. Delac, and S. Grgic. SCface - Surveillance Cameras Face Database. *Multimedia Tools and Application*, 51(3):863–879, 2011. XIX, XXIII, 4, 37, 45, 47, 54, 118, 129, 132, 134

[43] K. Grm, W. J. Scheirer, and V. Åătruc. Face hallucination using cascaded super-resolution and identity priors. *IEEE Transactions on Image Processing*, 29:2150–2165, 2020. 38, 51, 52

[44] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010. XIX, XXIII, 1, 19, 27, 28, 31, 119, 129

[45] G. Guo, L. Wen, and S. Yan. Face authentication with makeup changes. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):814–825, 2014. 95

[46] 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. 70, 82, 84, 85

[47] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 70, 71, 82, 86, 94, 106

[48] R. Hecht-Nielsen. Theory of the backpropagation neural network. In *International Joint Conference on Neural Networks*, pages 593–605, 1989. 127

[49] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, 2006. 120, 122, 130

[50] G. E. Hinton. A practical guide to training restricted Boltzmann machines. In *Neural Networks: Tricks of the Trade - Second Edition*, pages 599–619. Springer, 2012. 141, 144

[51] G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51, 2011. 18, 19, 21

[52] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computing*, 18(7):1527–1554, 2006. 130

[53] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 106

[54] C.-C. Hsu, C.-W. Lin, W.-T. Su, and G. Cheung. SiGAN: Siamese generative adversarial network for identity-preserving face hallucination. *IEEE Transactions on Image Processing*, 28(12):6225–6236, 2019. 38

[55] P. Hu and D. Ramanan. Finding tiny faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–959, 2017. 83

[56] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 76

[57] G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, 2014. 3, 12, 36, 59, 97, 112

[58] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 26

[59] K. A. Islam, D. Pérez, V. Hill, B. Schaeffer, R. Zimmerman, and J. Li. Seagrass detection in coastal water through deep capsule networks. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 320–331, 2018. 22

[60] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan. CapsuleGAN: Generative adversarial capsule network. In *European Conference on Computer Vision*, pages 526–535, 2018. 23

[61] M. Jian and K.-M. Lam. Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(11):1761–1772, 2015. 19, 110

[62] F. Juefei-Xu, D. K. Pal, K. Singh, and M. Savvides. A preliminary investigation on the sensitivity of COTS face recognition systems to forensic analyst-style face processing for occlusions. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–33, 2015. 57

[63] F. Juefei-Xu, E. Verma, P. Goel, A. Cherodian, and M. Savvides. Deepgender: Occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention. In *IEEE Computer Vision and Pattern Recognition Workshops*, pages 68–77, 2016. 118

[64] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa, and P. B. Sujit. DroneSURF: Benchmark dataset for drone-based face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2019. 6, 36, 37, 39, 46, 54

[65] B.-N. Kang, Y. Kim, B. Jun, and D. Kim. Attentional feature-pair relation networks for accurate face recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 5472–5481, 2019. 113

[66] H. Kazemi, F. Taherkhani, and N. M. Nasrabadi. Identity-aware deep face hallucination via adversarial face verification. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2019. 38

[67] 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. 59

[68] Y. Kim, W. Park, M.-C. Roh, and J. Shin. Groupface: Learning latent groups and constructing group-based representations for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5621–5630, 2020. 110, 113

[69] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 26, 106

[70] N. Kohli, D. Yadav, and A. Noore. Face Verification with Disguise Variations via Deep Disguise Recognizer. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–24, 2018. 68, 70

[71] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. 121

[72] S. Kumar, E. Yaghoubi, A. Das, B. Harish, and H. Proença. The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, Re-Identification and Search from Aerial Devices. *arXiv preprint arXiv:2004.02782*, 2020. 39

[73] V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa. Disguised faces in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 57

[74] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 53

[75] H. Larochelle and Y. Bengio. Classification using discriminative restricted Boltzmann machines. In *International Conference on Machine Learning*, pages 536–543, 2008. 130, 141, 144

[76] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin. Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, 10, 2009. 141

[77] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015. 119

[78] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 105–114, 2017. 20

[79] B. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna. Using Kinect for face recognition under varying poses, expressions, illumination and disguise. In *IEEE Workshop on Applications of Computer Vision*, pages 186–192, 2013. 60

[80] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5325–5334, 2015. 119

[81] P. Li, L. Prieto, D. Mery, and P. J. Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, 14(8):2000–2012, 2019. XXIII, 17, 19, 20, 31, 32, 33, 38, 45, 47, 48, 51, 52

[82] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017. 69, 70

[83] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, volume 2, 2016. 40, 47, 51

[84] X. Liu, S. Li, L. Kong, W. Xie, P. Jia, J. You, and B. Kumar. Feature-level frankenstein: Eliminating variations for discriminative recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 637–646, 2019. 105

[85] Y. Liu, H. Li, J. Yan, F. Wei, X. Wang, and X. Tang. Recurrent scale approximation for object detection in CNN. In *IEEE International Conference on Computer Vision*, pages 571–579, 2017. 71

[86] J. Lu, V. E. Liong, X. Zhou, and J. Zhou. Learning compact binary face descriptor for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):2041–2056, 2015. 57

[87] J. Lu, G. Wang, and P. Moulin. Human identity and gender recognition from gait sequences with arbitrary walking directions. *IEEE Transactions on Information Forensics and Security*, 9(1):51–61, 2014. 117

[88] Z. Lu, X. Jiang, and A. Kot. Deep coupled ResNet for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(4):526–530, 2018. 38, 110

[89] E. E. Maccoby and C. N. Jacklin. *The Psychology of Sex Differences*, volume 1. Stanford University Press, 1974. 117

[90] 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. 120, 122

[91] A. M. Martinez. The AR face database. *CVC Technical Report*, 1998. 1, 11, 59, 60

[92] A. M. Martínez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):748–763, 2002. 11, 97

[93] J. A. Martinez, P. C. Rutledge, and K. J. Sher. Fake ID ownership and heavy drinking in underage college students: Prospective findings. *Psychology of Addictive Behaviors*, 21(2):226, 2007. 138

[94] I. Masi, A. T. Trãn, T. Hassner, J. T. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, pages 579–596, 2016. 57

[95] F. V. Massoli, G. Amato, and F. Falchi. Cross-resolution learning for face recognition. *Image and Vision Computing*, 2020. doi: https://doi.org/10.1016/j.imavis.2020.103927. 4, 39, 44, 47, 48, 54

[96] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. 29, 49, 141

[97] Q. Meng, D. Catchpoole, D. Skillicom, and P. J. Kennedy. Relational autoencoder for feature extraction. In *International Joint Conference on Neural Networks*, pages 364–371, 2017. 120

[98] S. Menghani. Caught on camera: Couple abandon newborn at Delhi metro station. *News18*, 2016. XIII, 4

[99] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):707–711, 2002. 118

[100] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 86–94, 2017. 57

[101] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016. 57

[102] S. P. Mudunuri and S. Biswas. Low resolution face recognition across variations in pose and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):1034–1040, 2016. XV, 20, 28, 31, 33

[103] A. Nech and I. Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3406–3415, 2017. 59

[104] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. XIV, XXIII, 17, 19, 23, 24, 27, 29, 46, 52, 54

[105] A. Ng. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011. 120

[106] C.-B. Ng, Y.-H. Tay, and B.-M. Goi. A review of facial gender recognition. *Pattern Analysis and Applications*, 18(4):739–755, 2015. 117

[107] P. T. of India. CCTV images of suspect in woman IT employee murder released. 2016. XIII, 4

[108] M. O'Hare and S. Gupta. Young man caught posing as senior citizen to fly to US. *CNN Travel*, 2019. XIV, 10

[109] T. Ojala, M. Pietikäinen, and T. Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision*, pages 404–420, 2000. 1

[110] S. Ouyang, T. Hospedales, Y.-Z. 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. 19, 38, 110

[111] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. Overview of research on facial ageing using the FG-NET ageing database. *IET Biometrics*, 5(2):37–46, 2016. XX, 3, 138, 139

[112] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3):21–36, 2003. 20

[113] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, pages 41.1–41.12, 2015. 67, 70, 72, 119, 141, 144

[114] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 45

[115] S. V. Peri and A. Dhall. DisguiseNet : A Contrastive Approach for Disguised Face Verification in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–256, 2018. 68, 70

[116] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. An introduction to the good, the bad, and the ugly face recognition challenge problem. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 346–353, 2011. 76

[117] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):831–846, 2010. 76, 91

[118] P. Quintiliano and A. Santa-Rosa. Face recognition based on eigeneyes. *Pattern Recognition and Image Analysis*, 13(2):335–338, 2003. 11

[119] R. Raghavendra, N. Vetrekar, K. B. Raja, R. Gad, and C. Busch. Detecting disguise attacks on multi-spectral face recognition through spectral signatures. In *International Conference on Pattern Recognition*, pages 3371–3377, 2018. 11, 60

[120] R. Ramachandra and C. Busch. Presentation attack detection methods for face recognition systems: a comprehensive survey. *ACM Computing Surveys*, 50(1):1–37, 2017. 57

[121] N. Ramanathan, R. Chellappa, and A. R. Chowdhury. Facial similarity across age, disguise, illumination and pose. In *International Conference on Image Processing*, volume 3, pages 1999–2002, 2004. 60, 97

[122] R. Ranjan, C. D. Castillo, R. Chellappa, and R. Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 47, 51

[123] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 17–24, 2017. 72

[124] A. Rattani, C. Chen, and A. Ross. Evaluation of texture descriptors for automated gender estimation from fingerprints. In *European Conference on Computer Vision Workshops*, pages 764–777, 2014. 117

[125] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 66

[126] K. Ricanek Jr. and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2006. 140

[127] S. Rifai, G. Mesnil, P. Vincent, X. Muller, Y. Bengio, Y. Dauphin, and X. Glorot. Higher order contractive auto-encoder. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 645–660, 2011. 120

[128] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *International Conference on Machine Learning*, pages 833–840, 2011. 120

[129] M. S. Ryoo, K. Kim, and H. J. Yang. Extreme low resolution activity recognition with multi-Siamese embedding learning. In *AAAI Conference on Artificial Intelligence*, pages 7315–7322, 2018. 39

[130] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang. Privacy-preserving human activity recognition from extreme low resolution. In *AAAI Conference on Artificial Intelligence*, pages 4255–4262, 2017. 39

[131] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017. 21, 25, 29

[132] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *IEEE International Conference on Computer Vision*, pages 4501–4510, 2017. 20

[133] A. Sapkota and T. E. Boult. Large scale unconstrained open set face database. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–8, 2013. XXIII, 18, 19, 27, 31, 37, 45, 51, 54

[134] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 119

[135] J. Shi and C. Qi. From local geometry to global structure: Learning latent subspace for low-resolution face image recognition. *IEEE Signal Processing Letters*, 22(5):554–558, 2015. 47

[136] M. Singh, M. Chawla, R. Singh, M. Vatsa, and R. Chellappa. Disguised faces in the wild 2019. In *IEEE/CVF International Conference on Computer Vision Workshops*, pages 542–550, 2019. XIV, XVII, XXV, 10, 11, 87, 94, 96, 97, 104, 105, 107

[137] M. Singh, S. Nagpal, R. Singh, and M. Vatsa. Class representative autoencoder for low resolution multi-spectral gender classification. In *International Joint Conference on Neural Networks*, pages 1026–1033, 2017. XX, 127, 148

[138] M. Singh, S. Nagpal, R. Singh, and M. Vatsa. Dual directed capsule network for very low resolution image recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 340–349, 2019. XIV, 23, 34, 38, 46, 51, 52, 110

[139] M. Singh, S. Nagpal, R. Singh, and M. Vatsa. Derivenet for (very) low resolution image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. XV, 38, 55

[140] M. Singh, S. Nagpal, R. Singh, and M. Vatsa. Disguise resilient face verification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. XVIII, 98, 115

[141] 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*, 119:121–130, 2019. XX, 143, 148

[142] M. Singh, S. Nagpal, M. Vatsa, R. Singh, and A. Majumdar. Identity aware synthesis for cross resolution face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 479–488, 2018. 8, 17, 18, 20, 27, 31, 32, 38

[143] M. Singh, R. Singh, M. Vatsa, N. K. Ratha, and R. Chellappa. Recognizing disguised faces in the wild. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(2):97–108, 2019. XVI, XXV, 11, 58, 81, 83, 91, 94, 96, 97, 104, 105

[144] R. Singh, M. Vatsa, H. S. Bhatt, S. Bharadwaj, A. Noore, and S. S. Nooreyezdan. Plastic surgery: A new dimension to face recognition. *IEEE Transactions on Information Forensics and Security*, 5(3):441–448, 2010. 3

[145] R. Singh, M. Vatsa, and A. Noore. Face recognition with disguise and single gallery images. *Image and Vision Computing*, 27(3):245–257, 2009. 57, 60, 97

[146] 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. 1, 24

[147] E. Smirnov, A. Melnikov, S. Novoselov, E. Luckyanets, and G. Lavrentyeva. Doppelganger mining for face representation learning. In *IEEE International Conference on Computer Vision Workshops*, pages 1916–1923, 2017. 84, 85

[148] E. Smirnov, A. Melnikov, A. Oleinik, E. Ivanova, I. Kalinovskiy, and E. Luckyanets. Hard example mining with auxiliary embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 37–46, 2018. XVII, 67, 68, 69, 84, 85, 97, 105

[149] E. Smirnov, A. Oleinik, A. Lavrentev, E. Shulga, V. Galyuk, N. Garaev, M. Zakuanova, and A. Melnikov. Face representation learning using composite mini-batches. In *IEEE International Conference on Computer Vision Workshops*, pages 551–559, 2019. 107

[150] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu. Occlusion robust face recognition based on mask learning with pairwise differential Siamese network. In *IEEE/CVF International Conference on Computer Vision*, pages 773–782, 2019. 95

[151] A. Subramaniam, A. Narayanan Sridhar, and A. Mittal. Feature ensemble networks with re-ranking for recognizing disguised faces in the wild. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 97, 107

[152] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei. Circle loss: A unified perspective of pair similarity optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 113

[153] A. Suri, M. Vatsa, and R. Singh. A-LINK: Recognizing disguised faces via active learning based inter-domain knowledge. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2019. 97, 105

[154] S. Suri, A. Sankaran, M. Vatsa, and R. Singh. On matching faces with alterations due to plastic surgery and disguise. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2018. 76, 105

[155] S. Suri, A. Sankaran, M. Vatsa, and R. Singh. Improving face recognition performance using TeCS2 dictionary. *Pattern Recognition Letters*, 145:88–95, 2021. 105, 113

[156] C. Szegedy et al. Going deeper with convolutions. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015. 70

[157] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 119

[158] V. Talreja, F. Taherkhani, M. C. Valenti, and N. M. Nasrabadi. Attribute-guided coupled gan for cross-resolution face recognition. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2019. 51

[159] J. E. Tapia, C. A. Perez, and K. W. Bowyer. Gender classification from the same iris code used for recognition. *IEEE Transactions on Information Forensics and Security*, 11(8):1760–1770, 2016. 117

[160] L. Tran, X. Yin, and X. Liu. Representation learning by rotating your faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):3007–3021, 2018. 70

[161] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pages 1096–1103, 2008. 130

[162] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, 2010. 120, 121, 141, 144

[163] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004. 129, 135, 141

[164] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 41, 47, 51

[165] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 47, 48, 101, 113

[166] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. A comprehensive survey to face hallucination. *International Journal of Computer Vision*, 106(1):9–30, 2014. 20

[167] T. Y. Wang and A. Kumar. Recognizing human faces under disguise and makeup. In *IEEE International Conference on Identity, Security and Behavior Analysis*, 2016. 11, 60

[168] W. Wang, Y. Huang, Y. Wang, and L. Wang. Generalized autoencoder: A neural network framework for dimensionality reduction. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 496–503, 2014. 120

[169] X. Wang, K. Yu, C. Dong, and C. Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 20

[170] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. Studying very low resolution recognition using deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, 2016. 20, 27, 29, 30, 31, 38, 46, 51, 52

[171] Z. Wang, Z. Miao, Q. M. Jonathan Wu, Y. Wan, and Z. Tang. Low-resolution face recognition: A review. *The Visual Computer*, 30(4):359–386, 2014. 19, 110

[172] H. Wechsler, E. Jae, T. Nelson, and M. Kuo. Underage college students' drinking behavior, access to alcohol, and the influence of deterrence policies: Findings from the Harvard School of Public Health College Alcohol Study. *Journal of American College Health*, 50:223–236, 2002. 138

[173] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016. 42, 43, 69, 70

[174] C. Whitelam et al. IARPA Janus Benchmark-B Face Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 90–98, 2017. 113

[175] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Conference on Computer Vision and Pattern Recognition*, pages 529–534, 2011. 59, 112

[176] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2008. 97

[177] 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. 44, 82, 85, 86, 94, 110

[178] C. Xiang, L. Zhang, Y. Tang, W. Zou, and C. Xu. MS-CapsNet: A Novel Multi-Scale Capsule Network. *IEEE Signal Processing Letters*, 25(12):1850–1854, 2018. 23

[179] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):156–171, 2017. 57

[180] S. Yang, W. Deng, M. Wang, J. Du, and J. Hu. Orthogonality Loss: Learning Discriminative Representations for Face Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 113

[181] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. 69, 71, 82

[182] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley. Face super-resolution guided by facial component heatmaps. In *European Conference on Computer Vision*, pages 217–233, 2018. 110

[183] X. Yu, B. Fernando, R. Hartley, and F. Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 908–917, 2018. 110

[184] X. Yu, B. Fernando, R. Hartley, and F. Porikli. Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2926–2943, 2019. 38

[185] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang. Image super-resolution: The techniques, applications, and future. *Signal Processing*, 128:389 – 408, 2016. 38

[186] B. Zhang, L. Zhang, D. Zhang, and L. Shen. Directional binary code with application to PolyU Near-infrared Face Database. *Pattern Recognition Letters*, 31(14):2337–2344, 2010. 134

[187] K. Zhang, Y.-L. Chang, and W. Hsu. Deep disguised faces recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 32–36, 2018. XVII, 68, 69, 70, 97, 105

[188] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang. Super-identity convolutional neural network for face hallucination. In *European Conference on Computer Vision*, pages 183–198, 2018. 38, 51, 52

[189] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 69, 70, 71, 72

[190] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016. 47

[191] X. Zhang, Y. Fu, S. Jiang, L. Sigal, and G. Agam. Learning from synthetic data using a stacked multichannel autoencoder. In *International Conference on Machine Learning and Applications*, pages 461–464, 2015. 120

[192] K. Zhao, J. Xu, and M.-M. Cheng. Regularface: Deep face recognition via exclusive regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1144, 2019. 113

[193] S. Zhao, X. Gao, S. Li, and S. Ge. Low-resolution face recognition in the wild with mixed-domain distillation. In *IEEE International Conference on Multimedia Big Data*, pages 148–154, 2019. 39

[194] X. Zheng, Z. Wu, H. Meng, and L. Cai. Contrastive auto-encoder for phoneme recognition. In *International Conference on Acoustics, Speech and Signal Processing*, pages 2529–2533, 2014. 120, 121

[195] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li, and D. Wen. SFace: Sigmoid-Constrained Hypersphere Loss for Robust Face Recognition. *IEEE Transactions on Image Processing*, 30:2587–2598, 2021. 113

[196] 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. 49

[197] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *CoRR*, abs/1501.04690, 2015. 130, 141, 144

[198] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, and H. Ling. Vision meets drones: Past, present and future. *arXiv preprint arXiv:2001.06303*, 2020. 39

[199] S. Zhu, J. Qian, Y. Dong, and W. Wong. The Devil is in the Detail: Deep Feature Based Disguised Face Recognition Method. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 420–431, 2020. 105

[200] F. Zhuang, X. Cheng, P. Luo, S. J. Pan, and Q. He. Supervised representation learning: Transfer learning with deep autoencoders. In *International Joint Conference on Artificial Intelligence*, pages 4119–4125, 2015. 120

[201] W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21(1):327–340, 2012. 6, 17, 20, 35