Disguised Face Verification using Inverse Disguise Quality

Amlaan Kar[1], Maneet Singh[1], Mayank Vatsa[2], and Richa Singh[2]

[1]IIIT-Delhi, India; [2]IIT Jodhpur, India

Abstract. Research in face recognition has evolved over the past few decades. With initial research focusing heavily on constrained images, recent research has focused more on unconstrained images captured in-the-wild settings. Faces captured in unconstrained settings with disguise accessories persist to be a challenge for automated face verification. To this effect, this research proposes a novel deep learning framework for disguised face verification. A novel Inverse Disguise Quality metric is proposed for evaluating amount of disguise in the input image, which is utilized in likelihood ratio as a quality score for enhanced verification performance. The proposed framework is model-agnostic and can be applied in conjunction with existing state-of-the-art face verification models for obtaining improved performance. Experiments have been performed on the Disguised Faces in Wild (DFW) 2018 and DFW 2019 datasets, with three state-of-the-art deep learning models, where it demonstrates substantial improvement compared to the base model.

Keywords: Disguised Face Verification, Biometric Fusion

1 Introduction

Face recognition has witnessed substantial research interest with large number of applications, especially in social media, biometric authentication, social security, and enhanced user experience in the product domain. Several successful state-of-the-art face recognition models, including VGGFace [3], Residual Networks (ResNet) [4] and ArcFace [5], have been proposed in the literature. Research with these models has focused on covariates such as pose, illumination, expression, ageing, and heterogeneity, however, disguise variations have received limited research attention. Disguises can be considered as external "noise" which challenge the robustness of the face verification systems. Earlier research on disguised face recognition focused mostly on datasets prepared under constrained settings, thus failing to capture the real world scenario. Recently, the focus has shifted towards adapting face recognition to in-the-wild datasets like Disguised Faces in the Wild (DFW) 2018 [1] and 2019 [2], which incorporate both intentional and unintentional disguises ([18, 19]).

While deep neural networks such as VGGFace [3] and ResNet trained on face images have produced superlative performance on popularly used face recognition databases, their performance is subpar on disguise databases. Analysis of

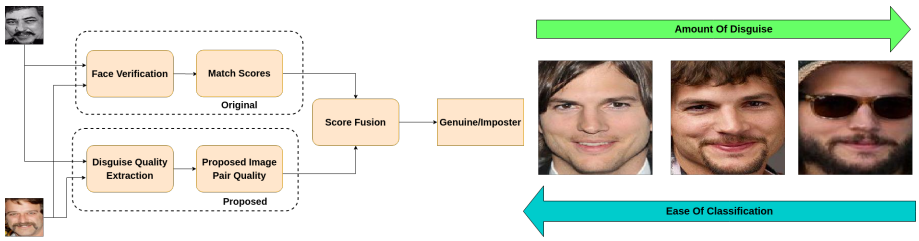


Fig. 1: (a) Proposed face verification algorithm where extracted image pair Inverse Disguise Quality is fused with match scores generated. (b) Inverse Disguise Quality denotes the biometric quality of a sample, i.e. on how easy it is to classify based on the "amount of disguise" present in the image.

the VGGFace model pre-trained on the VGGFace dataset, ResNet-50 model trained on the MS-Celeb-1M and VGGFace2 dataset [39] (referred to as VGGFace2), and the ResNet-100 model trained on MS1M-ArcFace dataset demonstrates verification accuracies of 33.76%, 66.97% and 65.51%, respectively, on the DFW 2018 dataset [1]. The poor performance of state-of-the-art face recognition models thus suggests a requirement for robust models invariant to disguise variations. Automated disguise face verification suffers from the challenge of both intentional and unintentional disguises. For example, concealing the identity, impersonating someone using glasses, moustache, beard, different hairstyles, scarfs or caps and makeup or even unintentionally changing the appearance as seen with hair growth or removal. These variations often result in reduced inter-class distance between subjects and increased intra-class variations, thus rendering the problem of disguised face recognition further challenging.

This research proposes a novel face verification framework for authenticating face images under disguise variations. The proposed framework utilizes the "amount of disguise" in an image to improve the performance of the face recognition algorithm in an attempt to make it more robust, secure, and usable in real-life scenarios. Fig. 1 shows how this "amount of disguise" is used as an image quality for this purpose. The contributions of this research are as follows:

- A novel face verification framework is proposed which utilizes an *Inverse Disguise Quality* metric for quantifying the biometric quality of a face image. The proposed framework is model-agnostic and can be applied in conjunction with existing state-of-art deep learning based face verification models for obtaining enhanced performance.
- Inverse Disguise Quality is derived by performing disguise detection on the face images. Semantic segmentation has been applied to recognize disguise and non-disguise patches, followed by a combination of the confidence scores for generating the quality metric. Further, the Inverse Disguise Quality has been fused with face verification match scores for improved performance.

– Experiments performed using state-of-the-art models pre-trained on large-scale face datasets demonstrate that the proposed framework yields improved performance as compared to the baseline models. For example, on the DFW 2018 dataset at 0.1% FAR, VGGFace [3], VGGFace2 [39] and ArcFace [5] models yield an overall increase of 63.41%, 38.11% and 35.07%, respectively.

2 Related Work

The proposed framework presents a novel technique for disguised face recognition. Disguise detection is performed on face images using semantic segmentation, followed by the estimation of Inverse Disguise Quality metric. The quality metric is fused with the match scores obtained via a face recognition model. This section presents the related work for the concepts of disguised face recognition, semantic segmentation, and likelihood ratio-based biometric score fusion.

2.1 Disguised Face Recognition

Initial research on disguised face recognition utilized datasets captured in constrained settings [20], [12], [21], [22], [23] and [24]. Chellappa et al. [12] studied the facial similarity for several variations including disguise by forming two eigenspaces from two halves of the face, one using the left half and other using the right half. From the test image, the optimally illuminated half face is chosen and projected into the eigenspace. Silva et al. [13] used the concept of eigeneyes to ensure that any change in facial features other than eyes does not affect the recognition performance. Singh et al. [14] used 2D log polar Gabor transform to extract phase features from faces, which were then divided into frames and matched using Hamming distance. Dhamecha et al. [15] classified the local facial regions of both visible and thermal face images into biometric and non-biometric classes. Yoon et al. [16] detected partially occluded faces using a SVM to characterize suspicious ATM users. From 2018, with the release of the Disguised Faces in the Wild (DFW) datasets [1] and [2], research started focusing more on unconstrained disguised face recognition. Smirnov et al. [17] proposed several ways to create auxiliary embeddings and used them to increase the number of potentially hard positive and negative examples. Zhang et al. [18] extracted features for generic faces using two networks. PCA based on the DFW 2018 dataset was applied to attempt a form of transfer learning. Deng et al. [19] applied the ArcFace [5] loss on the DFW 2018 [1] and DFW 2019 [2] datasets. They improved their generic implementation [5] by combining hard sample mining with the intra-loss and inter-loss. While the domain of disguised face recognition has attracted research focus in recent times, not-so-high performance of state-of-the-art algorithms as compared to non-disguised datasets demonstrate a need for further research.

2.2 Semantic Segmentation

Given an input image, semantic segmentation is the process of assigning a class label to each pixel for the purpose of object detection. Ciresan et al. [6] used a sliding widow to train a network, which would predict the class label of each pixel. This was done by providing a patch around that pixel as input. More recent approaches [7] utilized multiple resolutions of images to allow the use of localization and neighbourhood context at the same time. Long et al. [9] used fully convolutional networks to define a novel architecture that combines semantic and appearance information from different layers for segmentation. Ronneberger et al. [10] modified and extended this architecture and used an encoder-decoder model with skip connections to combine high resolution features with upsampled ones. Girshick et al. [40] combined several object detection predictions in images for semantic segmentation. He et al. [41] extended this by training a multi-branch network for bounding box recognition and object mask prediction. Semantic segmentation of faces has mostly focused on identifying different facial regions of the face image. Khan et al. [11] performed multi-class semantic segmentation on faces to separate various parts of the face using random forests. Jackson et al. [45] performed landmark localisation and then used it to guide semantic part segmentation. Kalayeh et al. [46] performed facial part parsing to transfer localisation properties to improve face attribute detection. Zhou et al. [47] combine fully-convolutional network, super-pixel information and CRF model to perform semantic segmentation. Lin et al. [48] used Mask RCNN [41] and FCN [9] branches for semantic labeling of the inner face and hair regions.

2.3 Likelihood Ratio and Quality in Biometric Fusion

At a conceptual level, the quality of any input (e.g. image or text) is a measure of the suitability of the input for automated analysis. For our task of face verification, a high quality sample is one which gets classified as genuine or imposter with relative ease as compared to a poor quality sample. Bharadwaj et al. [25] discussed several definitions and interpretations of biometric quality, and Singh et al. [35] performed a comprehensive survey of biometric fusion techniques. Bigun et al. [26] performed multimodal biometric fusion using a statistical framework that combined multiple verification scores along with corresponding quality metrics defined by human users. Fierrez-Aguilar et al. [27] used the quality of biometric samples as sample weights for training a SVM. Poh et al. [28] defined the quality of a biometric sample as how close the corresponding biometric sample is to the decision boundary that satisfies the equal error rate criterion. Nandakumar et al. [29] determined the quality of local regions in fingerprint and iris images to derive an overall quality of the match between each pair of template and query images. This quality was used to estimate a joint density between biometric match scores and their corresponding quality. Vatsa et al. [32] proposed applying Redundant Discrete Wavelet Transform (RDWT) to quantify distinguishing information present in an image. As far as the image quality in

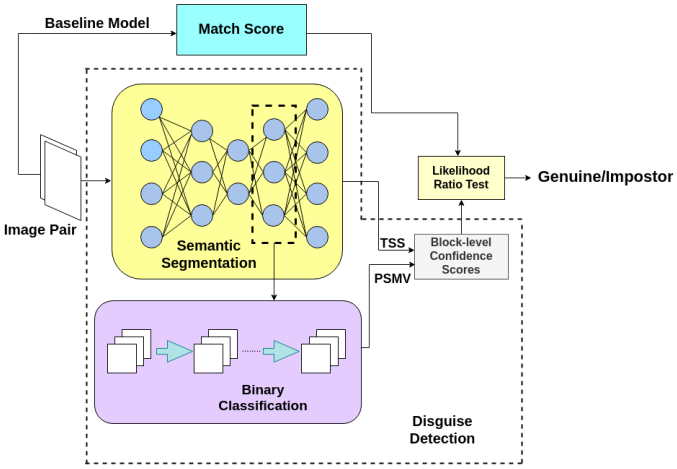


Fig. 2: The complete pipeline consisting of Disguise Detection and the usage of Inverse Disguise Quality in Face Verification.

face images is concerned, Subasic et al. [42] used 17 tests to determine face quality. These included image properties like resolution, brightness, aspect ratio, and sharpness, and facial properties like position and tilt of eyes and head. Gao et al. [43] used facial symmetry, lighting and pose to define face quality metrics. Zhang and Wang [44] extracted scale-invariant SIFT features from faces to define face asymmetry-based quality.

The proposed disguised face verification framework utilizes semantic segmentation for identifying disguised patches in a given input image. Based on the "amount of disguise", a novel Inverse Disguise Quality metric is calculated, followed by biometric fusion with the match scores obtained via the base network. The proposed framework is model-agnostic and can be applied in conjunction with existing deep learning based face verification models.

3 Proposed Disguised Face Verification Framework

The proposed framework (Fig. 2) consists of three components: (i) disguise detection on the input face image using thresholding on semantic segmentation to obtain block-wise semantic labelling, followed by the classification of block-wise learned features as disguise or non-disguise, (ii) computation of a quality metric based on the detected disguise which gives a quantitative measure of the amount of disguise in a face, and (iii) combination of the image quality into biometric pair quality for fusion with face verification scores obtained from a base network. The following subsections elaborate upon each component of the framework.

Fig. 3: Sample pairs of detected face and the corresponding semantic segmentation along with block-level annotation. We annotate each block in an image as disguise/non-disguise. Images are taken from the DFW 2018 [1] dataset.

3.1 Disguise Detection

In this research, the task of disguise detection involves predicting the regions of disguise or non-disguise in an image. The proposed pipeline for disguise detection broadly consists of three steps: (i) The input image is divided into 8x8 patches and semantic segmentation is performed on the image in order to classify each patch as disguise or non-disguise, (ii) A binary classifier is utilized for classifying the learned feature as disguise or non-disguise, and (iii) Weighted majority voting is performed by combining the 8 8 patches into 4 4 blocks for incorporating the neighbourhood information during disguise detection.

Step-1: Semantic Segmentation of Faces Semantic segmentation refers to the process of classifying each pixel in an image for the purpose of object detection. The task to be performed here is disguise detection in face images. In this research, semantic segmentation is applied to label blocks of the input image as *disguised* or *non-disguised* in a manner similar to [15]. The image is divided into 8 8 patches or blocks, and block-level labeling is performed. Thus, for an input image of dimension 224 224, 64 blocks are obtained. The U-net architecture [10, 33] is used for the said task. U-net is an encoder-decoder framework which performs semantic segmentation by downsampling an image in the encoder using convolution and up-sampling it using transpose convolutions to obtain pixel-level classification. It consists of skip connections between the encoder and decoder at equal levels of feature size to get precise locations by preventing information loss. Since U-net provides pixel-level labels, post training, thresholding is applied to transfer predicted pixel labels to their corresponding 28 28 block, to obtain the block-wise labels for disguise/non-disguise regions. In order to further strengthen the predictions at block-level, features are extracted from the trained model, followed by block-level classification into disguise/non-disguise.

Step-2: Binary Block Classification As demonstrated in Fig. 3, the semantic segmentation model described in the previous section provides a good

representation of disguised regions in an image. In order to further enhance the performance, block-level features are extracted from the trained U-net model for each image. As we separate the blocks of the image for classification, they become individual entities and retain no properties of being part of a larger image. This causes the neighbourhood of a block to no longer be explored, thus losing the structure of the image. On the other hand, semantic segmentation utilizes the whole image and not just different blocks. Thus, the features of semantic segmentation for a block also encode the neighbourhood of the block, therefore resulting in more descriptive and informative features.

The features for each block are provided as input to a binary block classifier, which classifies it as disguised or non-disguised. This module consists of six convolutional layers and two fully-connected layers. Further, since the number of non-disguised blocks are much more in number as compared to the disguised blocks, a weighted binary cross-entropy loss is applied to this model. We calculated the loss for weighted samples, i.e., individual weights are introduced for disguised samples, as shown in Equation 1.

$$C = \frac{1}{n} \sum_{x} [w_x(y \ln a + (1 \quad y) \ln(1 \quad a))] \tag{1}$$

where a, y are the predicted and ground-truth labels, respectively. x denotes the input and w_x is the weight corresponding to a given input. The above Equation allows us to increase the representation of disguised blocks in our training, as we are manually instructing the classification module to focus more on reducing the loss incurred due to the mis-classification of disguised blocks.

Step-3: Parent-Sibling Majority Voting In order to further incorporate the structure of the facial image, a parent-sibling majority voting is performed for obtaining the final disguise detection predictions. It is important to note that this component of the framework does not involve any training, and is applied directly on the predicted outputs of the binary block classifier. After getting predictions of all the 8 8 blocks, we classify each block on the basis of the predictions of blocks in its neighbourhood. Parent and siblings are the two types of neighbours considered here. Each block is seen as a part of a 4 4 block and each 4 4 block consists of four 8 8 blocks. For each block, we consider the corresponding 4 4 block as its parent. The other three 8 8 blocks corresponding to the 4 4 block are defined as the siblings of the block under consideration.

We follow majority voting among the four 8 8 blocks that constitute a 4 4 block to determine its label. Random selection of labels acts as a tie-breaker. A binary classifier, similar to the one described in the previous step, is used to classify 4 4 blocks as disguise/non-disguise. Once we get predictions of the 8 8 and 4 4 blocks, we find all the non-disguise confidence scores for 8 8 blocks. For all the blocks classified with confidence scores below a threshold, we follow a weighted majority voting between the disguise and non-disguise confidence scores of the siblings and parent of that block. The 4 4 block is given a weight of 0.7, while the siblings are given weights of 0.1 each. This weight assignment

has been done to take into account the fact that the 4 4 block best represents the neighbourhood of a given 8 8 block. The proposed voting mechanism helps low confidence samples to be represented via a more confident label obtained from its neighbourhood. Mathematically, this is denoted as:

$$Class = \arg\max_{C} \left\{ \left(w \quad D^{4\times4} + \left(\frac{1 \quad w}{3} \right) \quad \sum_{n} \left[D_n^{8\times8} \right] \right), \left(w \quad N^{4\times4} + \left(\frac{1 \quad w}{3} \right) \quad \sum_{n} \left[N_n^{8\times8} \right] \right) \right\}$$

where C, w, n, D and N denote the set of classes, hierarchy level weight, 8 8 neighbours, disguise and non-disguise confidence score, respectively.

3.2 Inverse Disguise Quality for Face Verification

The result obtained from the disguise detection module are utilized to compute a novel *Inverse Disguise Quality* metric for quantifying the quality of the input image. The inverse disguise quality score is further fused with the match scores obtained via a base face verification model for obtaining improved performance.

Inverse Disguise Quality The quality of biometric samples has a significant impact on the accuracy of a matcher. Poor quality biometric samples often lead to incorrect matching results since the features extracted from them are not reliable. Therefore, assigning weights to the predicted output of a face verification model based on the quality of the input sample can improve the overall recognition performance. In this research, a novel Inverse Disguise Quality metric is used to quantify the "amount of disguise" in an image. As disguise proves to be an obstruction in face recognition or verification, the amount of disguise in an image is an ideal metric to quantify whether the input is of good biometric quality or not. Hence, we define the Inverse Disguise Quality of an image as the sum of the non-disguised confidence scores of each block of the image.

$$\mathsf{D} = \sum_{i=1}^{64} [(Non \quad DisguiseConfidence)_i] \tag{2}$$

where, D denotes the inverse disguise quality. If the amount of non-disguised blocks are higher, the inverse disguise quality metric will also be higher, thus suggesting more confidence in the feature extraction by the base model.

Inverse Disguise Quality based Likelihood Ratio Test The Inverse Disguise Quality is fused with the likelihood ratio for enhanced disguised face verification. The likelihood ratio test utilizes the match score densities obtained for the genuine and impostor classes. In this research, we have used a Gaussian distribution [30] to estimate the match score densities. Let s, $f_{gen}(s_{gen})$ and

$f_{imp}(s_{imp})$ denote the vector of match scores and the conditional densities of the genuine and impostor match scores, respectively. Traditionally, let Ψ be a statistical test for testing. For an input I_s: H_0: I_s corresponds to an impostor and H_1: I_s corresponds to a genuine user. As per the Neyman-Pearson theorem [36], for testing H_0 against H_1, there exists a test Ψ and a constant η such that:

$$P(\Psi(s) = 1 j H_0) = \alpha \tag{3}$$

$$\Psi(I_s) = \begin{cases} 1, & \frac{f_{\text{gen}}(I_s)}{f_{\text{imp}}(I_s)} \quad \eta \\ 0, & \frac{f_{\text{gen}}(I_s)}{f_{\text{imp}}(I_s)} \quad \eta \end{cases} \tag{4}$$

If a test satisfies these two equations for some η, then it is the most powerful test for testing H_0 against H_1 at level α. If the false accept rate (FAR) α is fixed, the optimal test for deciding which class I_s belongs to is the likelihood ratio test $\Psi(I_s)$ if $f_{gen}(I_s)$ and $f_{imp}(I_s)$ are estimated properly. In other words, any data point with likelihood ratio η will be classified as belonging to the genuine class. All other data samples will be classified as impostor.

In the proposed framework, the Inverse Disguise Quality metric is incorporated into the likelihood ratio test for face verification. For this purpose, we have used two quality metrics for image pairs: (i) Average Image Pair Quality (AIPQ), where we take the average of the inverse disguise image qualities of the image pair; (ii) Normalized Image Pair Quality (NIPQ), where the image qualities are normalized in the range of [0,1]. The quality metric is incorporated into the likelihood ratio test by multiplying each element of the log-likelihood ratio vector with the corresponding image pair quality, taking from Poh and Bengio [28], where the independent match scores are combined with the corresponding sample quality in the authentication phase. Mathematically, it is expressed as:

$$\Psi(I_s) = \begin{cases} 1, & Q(I) \quad \ln \frac{f_{\text{gen}}(I_s)}{f_{\text{imp}}(I_s)} \quad \eta_1 \\ 0, & Q(I) \quad \ln \frac{f_{\text{gen}}(I_s)}{f_{\text{imp}}(I_s)} < \eta_1 \end{cases} \tag{5}$$

where $Q(I)$ denotes AIPQ or NIPQ of the input image pair I. Multiplying the log-likelihood ratios helps in enhanced face verification performance since the inherent concept of quality of an image is that a poor quality sample will be difficult to classify as genuine or impostor. By multiplying log-likelihood ratios with the corresponding image pair quality, we explicitly weigh the resulting log-likelihood ratios with their corresponding image pair qualities. For example, if face verification is performed using NIPQ, we observe that multiplication of the log-likelihood ratios and NIPQ leads to downscaling of the LRs as NIPQ lies between 0 and 1. Multiplication helps us achieve the purpose of trusting good quality samples more and poor quality samples less, as pairs with low quality are downscaled to an extent greater than that of good quality pairs.

4 Dataset and Implementation Details

The proposed framework has been evaluated for disguised face verification on two recent challenging datasets: DFW 2018 and DFW 2019 datasets. Details regarding the datasets, protocols, and implementation details are as follows.

4.1 Datasets

DFW 2018 Dataset was released as part of the DFW Workshop in CVPR 2018 [1]. The dataset contains 11,157 face images belonging to 1000 subjects having unconstrained disguise variations. Out of these 1000 subjects, 400 subjects belong to the training set, and 600 belong to the testing set. Each subject contains at least 5 and at most 26 face images of normal, validation, disguise and impersonator types.The dataset has been released with three protocols for evaluation. For all protocols, face verification is to be performed between a gallery and a probe image for classification as genuine or imposter.
DFW 2019 Dataset: It was released as part of the DFW Workshop in ICCV 2019 [2]. The dataset was released as a test set only, while encouraging researchers to utilize the DFW 2018 dataset as the training and validation set. The DFW 2019 dataset contains images pertaining to 600 subjects. Further, the dataset contains variations due to bridal make-up, and out of the 600 subjects, 250 subjects demonstrate variations due to plastic surgery. As compared to DFW 2018, an additional protocol related to plastic surgery has also been added. The DFW 2019 dataset has been released with four protocols for evaluation. For all four protocols, face verification is to be performed between a gallery and a probe image for classification as genuine or imposter. Detailed description of each protocol [1, 2] is given as:

- **Protocol-1 (Impersonation)** is defined for the purpose of differentiating between genuine users and impersonators. A genuine pair of gallery and probe images consists of same subject images while an imposter pair consists of pairs of impersonator images with the other types of images for a subject. This protocol is present in both DFW 2018 and DFW 2019 datasets.
- **Protocol-2 (Obfuscation)** evaluates the ability of a face verification framework to perform accurately even with obfuscated face images. Genuine pairs include disguised images of a subject paired with normal, validation and other disguised images of the same subject. The imposter set is created by combining cross-subject normal, validation, and disguised images. This protocol is present in both DFW 2018 and DFW 2019 datasets.
- **Protocol-3 (Plastic Surgery)** evaluates the ability of a face verification framework to identify faces despite changes in them due to plastic surgery. The genuine pairs are made of same subject pre-surgery and post-surgery images while the imposter set contains cross-subject pre-surgery and post-surgery images. This protocol is present only in the DFW 2019 dataset.
- **Protocol-4 (Overall)** is used to evaluate the performance of any face recognition algorithm on the entire DFW dataset. The genuine and imposter sets

created in the above mentioned protocols are combined to generate the data for this protocol. The overall protocol is Protocol 3 in DFW 2018 and is formed by combining Protocols 1 and 2. In DFW 2019, it is formed by combining all three protocols described above.

4.2 Implementation Details

For both the datasets, face detection is performed using the face coordinates provided by the authors. Since there does not exist any dataset with annotated disguise patches, manual annotation is performed on the training images for obtaining the ground-truth labels for the disguise detection module. The face images are resized to 224 224, and divided into blocks of 8 8. Each block is annotated as *disguised* or *non-disguised*, where a disguised block is defined as a block with at least 25% coverage of the disguise accessory. The annotation procedure is performed in an absolute manner for external objects obstructing the face like caps, hats, scarves, glasses, etc. Any such external accessory is marked as disguise. Factors like facial hair and hair colour are considered relative to the normal image. Every block in a normal image which is not affected by an external accessory is considered as non-disguised. For validation and disguised images, blocks containing facial hair and hair colour are annotated with respect to the normal image. Impersonator image blocks containing facial hair are marked as disguise if facial hair is present in the corresponding region in the normal image as well. The annotated images are passed through a U-Net for semantic segmentation, as described in Section 3.1. The model is trained using the Adam optimizer and the dice coefficient loss function with a learning rate of $1e$ 3. Features of the last convolutional layer are separated into 8 8 blocks for binary classification, which is done using a classification network consisting of six convolutional layers and two fc-layers. As described in Section 3.1, training is done using weighted binary cross entropy loss. The weights of the disguised blocks for the separate 8 8 and 4 4 networks are selected as 7 and 21 respectively for the best average between block-level accuracies of disguised and non-disguised blocks. Training was done using the Adam optimizer at a learning rate of $1e$ 5.

5 Experimental Results and Analysis

The experiments for disguise detection are performed on the DFW 2018 dataset, and face verification experiments on the DFW 2018 and DFW 2019 datasets. For face verification, experiments are performed on all the protocols of both the datasets. The efficacy of the proposed framework is demonstrated by applying it on existing pre-trained models. Three base models are used: (i) VGGFace model pre-trained on the VGGFace dataset [3], (ii) ResNet-50 model trained on the MS-Celeb-1M and VGGFace2 dataset [39] (referred to as VGGFace2), and (iii) ArcFace model (ResNet-100) trained on the MS1M dataset [5].

Table 1: Block-level disguise detection accuracy (%) on the DFW 2018 dataset. Results have been computed for the thresholding on semantic segmentation (TSS), binary classification (BC) and parent-sibling majority voting (PSMV).

Technique	**TSS**	**BC (8 X 8)**	**BC (4 X 4)**	**PSMV**
Disguised Blocks (%)	56.43%	68.66%	76.45%	69.32%
Non-Disguised Blocks (%)	89.66%	77.47%	66.54%	77.67%

Table 2: Verification accuracy (%) for the impersonation (P-1), obfuscation (P-2), and overall (P-3) protocols of the DFW 2018 dataset. The proposed technique with NIPQ scores demonstrates improved performance as compared to the baseline model across different False Acceptance Rates (FAR).

Prtcl.	**Algo.**	**GAR@1% FAR**			**GAR@0.1% FAR**		
		VGGFace	**VGGFace2**	**Arcface**	**VGGFace**	**VGGFace2**	**Arcface**
P-1	**Base**	52.77%	80.17%	87.22%	27.05%	48.23%	55.79%
	AIPQ	80.00%	92.60%	91.93%	71.43%	86.05%	77.64%
	NIPQ	95.24%	97.98%	**98.46%**	93.10%	**96.97%**	95.12%
P-2	**Base**	31.52%	66.32%	64.10%	15.72%	43.79%	45.48%
	AIPQ	54.65%	77.81%	75.56%	37.92%	66.75%	61.58%
	NIPQ	89.26%	89.85%	**90.52%**	79.82%	**80.89%**	80.23%
P-3	**Base**	33.76%	66.97%	65.51%	17.74%	44.05%	47.51%
	AIPQ	56.39%	78.51%	76.03%	40.28%	67.27%	62.81%
	NIPQ	91.47%	91.92%	**92.43%**	81.14%	82.16%	**82.58%**

5.1 Disguise Detection Results

Table 1 summarizes the block classification accuracies on the DFW 2018 dataset obtained by thresholding on semantic segmentation (TSS), binary classification (BC) and parent-sibling majority voting (PSMV). BC improves the disguised block accuracy by 12.23%, as compared to the traditional semantic segmentation. While the non-disguised block accuracy comes down by 12.19%, this is because BC removes the bias present due to TSS. PSMV improves disguised and non-disguised block accuracies by 0.66% and 0.20%, respectively and achieves the best performance for disguise detection.

5.2 Face Verification Performance

The proposed framework has been evaluated on the DFW 2018 and DFW 2019 test datasets. Fig. 4 presents the Receiver-Operator Characteristics (ROC) curves for the respective best model on the protocols of the DFW 2018 and the DFW 2019 datasets. Table 2 shows the verification accuracies of the baseline models and the proposed frameworks for the protocols of the DFW 2018 dataset. Table 3 shows the same for the protocols of the DFW 2019 dataset.

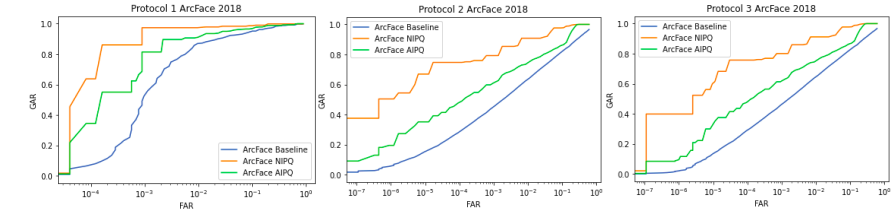
Table 3: Verification accuracy (%) for the impersonation (P-1), obfuscation (P-2), plastic surgery (P-3), and overall (P-4) protocols of the DFW 2019 dataset. The proposed technique with NIPQ scores demonstrates improved performance as compared to the baseline model across different FARs.

Prtcl.	Algo.	GAR@0.01% FAR			GAR@0.1% FAR		
		VGGFace	VGGFace2	Arcface	VGGFace	VGGFace2	Arcface
P-1	Base	14.80%	27.20%	7.20%	26.00%	57.60%	47.20%
	AIPQ	48.00%	63.60%	9.20%	86.80%	72.80%	65.60%
	NIPQ	78.40%	**92.80%**	24.80%	86.80%	**98.00%**	88.80%
P-2	Base	3.71%	35.55%	12.33%	10.05%	55.37%	25.34%
	AIPQ	22.92%	60.85%	26.55%	26.88%	75.61%	42.75%
	NIPQ	73.68%	**90.22%**	55.47%	73.68%	**93.57%**	74.97%
P-3	Base	7.20%	35.60%	34.40%	14.00%	60.40%	54.40%
	AIPQ	14.80%	40.40%	38.00%	18.80%	68.00%	59.60%
	NIPQ	50.80%	**72.80%**	59.60%	50.80%	**82.80%**	72.00%
P-4	Base	3.11%	34.12%	14.12%	9.52%	54.70%	27.16%
	AIPQ	25.48%	62.21%	28.48%	27.41%	73.03%	44.59%
	NIPQ	74.81%	**90.59%**	57.03%	74.81%	**93.80%**	75.94%

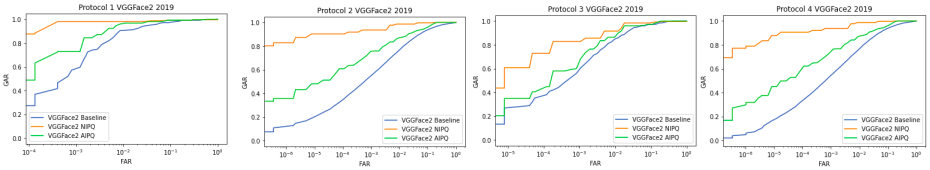
Analysis with Different Base Models: On the DFW 2018 dataset, the ArcFace model gives the best baseline verification accuracies. On application of NIPQ, it is observed from Table 2 that the performance of VGGFace and VGGFace2 is comparable to that of ArcFace. At 0.1% FAR, the performance of the three models improve by 63.41%, 38.11% and 35.07% respectively. VGGFace2 easily outperforms the other two models on DFW 2019 dataset for both the baseline and NIPQ. At 0.1% FAR, the performance of the three models improve by 65.29%, 39.10% and 48.78% respectively. The improved performance after applying the proposed framework motivates its usage with different models.

Protocol-specific Performance: All three base models show substantial improvements on all the protocols in both datasets. On DFW 2018 dataset, VGGFace2 is the best performing model for Protocols 1 and 2 with NIPQ showing improvements of 48.74% and 37.10% respectively over the baseline at 0.1% FAR. ArcFace is the best performing model on the overall protocol with a corresponding improvement of 35.07%. On the DFW 2019 dataset, Table 3 shows that VGGFace2 is the best performing model on all four protocols, with NIPQ improving over the baseline by 40.40%, 38.20%, 22.40% and 39.10% respectively at 0.1% FAR.

Effect of Quality Metric: As described in Section 3.2, we have shown results for AIPQ and NIPQ, where AIPQ is the average of the Inverse Disguise Qualities of an image pair while NIPQ is obtained by normalizing AIPQ to lie between 0 and 1. ArcFace, the best performing model on DFW 2018 dataset, shows an improvement of 10.92%, 14.14% and 14.89% when NIPQ is used instead of



(a) P-1 (b) P-2 (c) P-3



(d) P-1 (e) P-2 (f) P-3 (g) P-4

Fig. 4: ROC curves for the best models on all protocols of the DFW 2018 dataset ((a)-(c)) [1] and the DFW 2019 dataset ((d)-(g)) [2].

AIPQ on Protocols 1, 2 and 3 of DFW 2018 dataset at 0.1% FAR respectively. The corresponding results for VGGFace2 on the DFW 2019 dataset are 25.20%, 17.96%, 14.80% and 10.77% respectively. The results motivate the inclusion of NIPQ in the proposed framework.

6 Conclusion

This research proposes a novel framework for disguised face verification incorporating the proposed Inverse Disguise Quality metric. The framework is model-agnostic and can be applied in conjunction with existing deep learning based face verification models. Disguise detection is performed on the input face image using a combination of semantic segmentation and binary classification models. Based on the predictions, the Inverse Disguise Quality Metric is computed which provides an estimate of the image's quality. The proposed metric has been incorporated into the likelihood-ratio based verification process for obtaining enhanced performance. Experiments have been performed on the recently released Disguised Faces in the Wild (DFW) 2018 and DFW 2019 datasets. Analysis is drawn using three state-of-the-art models: (i) VGGFace, (ii) VGGFace2, and (iii) ArcFace, where, substantial improvement is observed in the verification performance upon applying the proposed framework.

References

1. 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, vol. 1, no. 2, pp. 97–108, 2019.
2. M. Singh, M. Chawla, R. Singh, M. Vatsa, and R. Chellappa. Disguised faces in the wild 2019. In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.
3. O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In British Machine Vision Conference, volume 1, page 6, 2015.
4. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.
5. J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4690–4699, 2019.
6. D.C. Ciresan, L.M. Gambardella, A. Giusti, J. Schmidhuber, Deep neural networks segment neuronal membranes in electron microscopy images, In Neural Information Processing Systems, pp. 2852–2860, 2012.
7. M. Seyedhosseini, M. Sajjadi, T. Tasdizen, Image segmentation with cascaded hierarchical models and logistic disjunctive normal networks, In IEEE International Conference on Computer Vision (ICCV), pp. 2168–2175, 2013.
8. B. Hariharan, P. Arbelez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 447-456. 2015.
9. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440, 2015.
10. O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer Assisted Intervention, pages 234–241. Springer, 2015.
11. Khalil Khan, Massimo Mauro, and Riccardo Leonardi, "Multi-class semantic segmentation of faces," in IEEE International Conference on Image Processing (ICIP), pp. 827–831, 2015.
12. N. Ramanathan, A. R. Chowdhury, and R. Chellappa, "Facial similarity across age, disguise, illumination and pose," Proceedings of International Conference on Image Processing, Vol. 3, pp. 1999 - 2002, 2004.
13. P. Q. Silva and A. N. C. Santa Rosa, "Face recognition based on eigeneyes," Pattern Recognition and Image Analysis, Vol. 13, No. 2, pp. 335 - 338, 2003.
14. R. Singh, M. Vatsa, and A. Noore, "Face recognition with disguise and single gallery images", Image and Vision Computing, vol. 27, no. 3, pp. 245–257, 2009.
15. T.I. Dhamecha, A. Nigam, R. Singh, and M. Vatsa, Disguise detection and face recognition in visible and thermal spectrums. In The International Conference on Biometrics, pp. 1–8, 2013.
16. J. Kim, Y. Sung, S. M. Yoon, and B. G. Park, "A new video surveillance system employing occluded face detection," in International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 65–68, Springer, 2005.
17. E. Smirnov, A. Melnikov, A. Oleinik, E. Ivanova, I. Kalinovskiy, and E. Luckyanets, "Hard example mining with auxiliary embeddings," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 37–46, 2018.
18. K. Zhang, Y.-L. Chang, and W. Hsu, "Deep disguised faces recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 32–36, 2018.

19. J. Deng and S. Zafeririou 2019. Arcface for disguised face recognition. In IEEE International Conference on Computer Vision Workshop on Disguised Faces in the Wild, pp. 0-0, 2019.
20. A. M. Martinez, "The AR face database," CVC Technical Report24, 1998.
21. B. Y. 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, pp. 186–192, 2013.
22. T. Y. Wang and A. Kumar, "Recognizing human faces under disguise and makeup," in IEEE International Conference on Identity, Security and Behavior Analysis, pp. 1–7, 2016.
23. 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, pp. 3371–3377, 2018.
24. A. Singh, D. Patil, M. Reddy, and S. Omkar, "Disguised face identification (dfi) with facial keypoints using spatial fusion convolutional network,"in Proceedings of the IEEE International Conference on Computer Vision, pp. 1648–1655, 2017.
25. S. Bharadwaj, M. Vatsa, and R. Singh, Biometric Quality: A Review of Fingerprint, Iris, and Face, EURASIP Journal of Image and Video Processing, 2014(1), 34
26. J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, Multimodal biometric authentication using quality signals in mobile communications, In 12th International Conference on Image Analysis and Processing, pp. 2-11, 2003.
27. J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Discriminative Multimodal Biometric Authentication Based on Quality Measures," Pattern Recognition, vol. 38, no. 5, pp. 777-779, 2005.
28. N. Poh, S. Bengio, Improving fusion with margin-derived confidence in biometric authentication tasks, in: International Conference on Audio And Video-Based Biometric Person Authentication, pp. 474–483, 2005.
29. K. Nandakumar, Y. Chen, A. K. Jain, S. C. Dass, Quality-based score level fusion in multibiometric systems, in: International Conference on Pattern Recognition, pp. 473–476, 2006.
30. K. Nandakumar, Y. Chen, S. C. Dass, A. Jain, Likelihood ratio-based biometric score fusion, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2), pp. 342-347, 2008.
31. N. Poh, J. Kittler, T. Bourlai, Improving biometric device interoperability by likelihood ratio-based quality dependent score normalization, In 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems, pp. 1-5. IEEE, 2007.
32. M. Vatsa, R. Singh, A. Noore, Integrating image quality in 2ν-SVM biometric match score fusion, International Journal of Neural Systems 17 (05) (2007) 343–351.
33. https://github.com/ZFTurbo/ZF_UNET_224_Pretrained_Model
34. K. Dharavath, F. A. Talukdar, and R. H. Laskar, "Improving face recognition rate with image preprocessing," Indian Journal of Science and Technology, vol. 7, no. 8, pp. 1170–1175, 2014.
35. M. Singh, R. Singh, A. Ross, A Comprehensive Overview of Biometric Fusion, Information Fusion, Volume 52, Pages 187-205, 2019.
36. E.L. Lehmann and J.P. Romano, Testing Statistical Hypotheses. Springer Science & Business Media, 2006.
37. Iacopo Masi and Anh Tuãn Trãn and Tal Hassner and Jatuporn Toy Leksut and Gérard Medioni, "Do we really need to collect millions of faces for effective face

recognition?" In European Conference on Computer Vision, pp. 579–596, Springer, 2016.
38. 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, pp. 87–102, Springer, 2016. 19
39. 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 and Gesture Recognition, pp. 67–74, 2018.
40. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Computer Vision and Pattern Recognition, pp. 580-587, 2014.
41. K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN, In International Conference on Computer Vision (ICCV), pp. 2961-2969, 2017.
42. M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec. Face image validation system. In International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 30–33, 2005.
43. X. Gao, S. Li, R. Liu, P. Zhang, Standardization of face image sample quality, in Proceedings of Advances in Biometrics, pp. 242–251, 2007.
44. G. Zhang, Y. Wang, Asymmetry-based quality assessment of face images, in Advances in Visual Computing, Las Vegas, vol. 5876, Springer, Berlin Heidelberg, pp. 499–508, 2009.
45. A.S. Jackson, M. Valstar, and G. Tzimiropoulos, A CNN cascade for landmark guided semantic part segmentation, In European Conference on Computer Vision (pp. 143-155), Springer, Cham, 2016.
46. Mahdi M. Kalayeh, Boqing Gong, Mubarak Shah, Improving facial attribute prediction using semantic segmentation, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6942-6950, 2017.
47. L. Zhou, Z. Liu, X. He, Face Parsing via a Fully-Convolutional Continuous CRF Neural Network, arXiv-1708, 2017.
48. J. Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, Lu Yuan, Face parsing with RoI tanh-warping, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5654-5663, 2019.