Evading Face Recognition via Partial Tampering of Faces

Puspita Majumdar, Akshay Agarwal, Richa Singh, and Mayank Vatsa
IIIT-Delhi, India
{pushpitam, akshaya, rsingh, and mayank}@iiitd.ac.in

Abstract

Advancements in machine learning and deep learning techniques have led to the development of sophisticated and accurate face recognition systems. However, for the past few years, researchers are exploring the vulnerabilities of these systems towards digital attacks. Creation of digitally altered images has become an easy task with the availability of various image editing tools and mobile application such as Snapchat. Morphing based digital attacks are used to elude and gain the identity of legitimate users by fooling the deep networks. In this research, partial face tampering attack is proposed, where facial regions are replaced or morphed to generate tampered samples. Face verification experiments performed using two state-of-the-art face recognition systems, VGG-Face and OpenFace on the CMU- MultiPIE dataset indicates the vulnerability of these systems towards the attack. Further, a Partial Face Tampering Detection (PFTD) network is proposed for the detection of the proposed attack. The network captures the inconsistencies among the original and tampered images by combining the raw and high-frequency information of the input images for the detection of tampered images. The proposed network surpasses the performance of the existing baseline deep neural networks for tampered image detection.

1. Introduction

Face recognition systems are used in a wide range of applications ranging from e-payments, automatic border control access through e-pass and surveillance. The advancement in machine learning and deep learning techniques with the wide availability of training data have led to the development of sophisticated deep learning algorithms for face recognition [4, 26, 30, 40]. However, the vulnerability of deep face recognition systems towards digital attacks is a major concern. With the advancement of sophisticated and easy to use image editing tools and mobile applications such as *Snapchat*, creating digitally altered images has become an easy task.

Digital attacks are of various types including morphing based attacks, retouching based attacks, and adversarial attacks. In morphing based attacks, a new face image is generated using the information available from multiple source face images of different subjects to elude own identity or gain the identity of others. In the literature, researchers have shown the vulnerability of face recognition systems towards morphing based digital attacks [1, 10, 18, 21, 22, 33, 34]. However, due to morphing, the visual appearance of the images changes to some extent. Retouching on the other hand affects the performance of recognition systems by changing the geometric properties of the face image which in turn changes the visual appearance of the images [5, 6]. In learning based adversarial attacks, adversaries in the form of visually imperceptible noise are added to the input images to deteriorate the performance of deep networks [7, 14, 15, 29, 38]. However, such attacks require knowledge of the model to attack.

Figure 1. Guess which of the images in the second and third row are original or tampered? Hint: Top row contains the source images used to create the tampered images.

Figure 1 shows some samples of digitally altered images generated by partial replacement and morphing of facial re-

gions. Figure 1 illustrates that it is quite difficult to differentiate among the original and tampered images. Therefore, the images of the same subjects can be easily identified by humans due to the similarity in the visual appearance of the original and tampered images. However, it is asserted that morphing and replacement of specific parts of human face with other subjects could present new challenges to face recognition systems. This may exploit vulnerabilities in the learned parameters if certain parts of the face are weighted over the others.

This research focuses on answering the question: "*are existing deep face recognition systems robust towards minuscule changes in facial regions?*" In this research, a partial face tampering attack is proposed by partially replacing and morphing of facial regions. The proposed attack does not require the knowledge of the system to attack, and the visual appearance of the tampered images remains unchanged. The first aim is to analyze the robustness of existing deep face recognition systems towards minute changes in facial regions imperceptible to human eye. Secondly, a novel tampered image detection network termed as *Partial Face Tampering Detection (PFTD)* is proposed for detecting the proposed attack. The network uses a combination of RGB image and high pass filtered version of the input image to detect the tampered images. The main contributions of this research are summarized below:

- Generation of partial face tampered samples using replacement and morphing of facial parts;
- Performance analysis of OpenFace [4] and VGG-Face [30] models through face verification experiments;
- Proposing a Partial Face Tampering Detection (PFTD) network for the detection of the proposed partial face tampering attack.
- Experiments for detection of unseen digital attacks are also performed to showcase the effectiveness of PFTD network.

The remaining paper is organized as follows: Section 2 presents the related work, Section 3 discusses the proposed partial face tampering attack with its effect on OpenFace and VGG-Face. Section 4 gives the details of the proposed Partial Face Tampering Detection network with results and analysis. Finally, Section 5 concludes the paper.

2. Related Work

In the literature, vulnerability of deep learning algorithms towards adversarial attacks [3, 7, 28, 29, 38] and deep face recognition systems towards face morphing or swapping [1, 9, 34] are highlighted by several researchers. In 2017, Agarwal et al. [1] have shown the effect of morphed face images on Commercial-Off-The-Shelf System (COTS) by creating a novel SWAPPED-Digital Attack Video Face Database using Snapchat. Further, the authors proposed a weighted local magnitude patterns with Support Vector Machine (SVM) classifier for the detection of morph faces. Scherhag et al. [34] investigated the vulnerabilities of biometric systems towards morphed face attacks. Other work on the detection of morph faces includes [24, 31]. Raghavendra et al. [31] proposed a feature level fusion approach of two pre-trained CNN networks for the detection of digital and print-scanned morphed face images. Recently, Ferrara et al. [11] have shown the effect of morphing on COTS and proposed a technique to demorph the morphed face image.

Apart from the analysis and detection of morphing based attacks, several algorithms have been proposed for the detection of adversarial attacks. Goswami et al. [15] proposed a selective dropout approach to detect adversarial samples. Lu et al. [25] proposed a Radial Basis Function SVM classifier to detect adversarial samples. Metzen et al. [27] proposed to augment a subnetwork trained for classifying adversarial samples to a targeted network. Goel et al. [12] have implemented the adversarial examples generation and detection algorithms and prepared a toolbox called Smartbox. Other works for the detection of adversarial samples include [2, 17, 19, 23]. A detailed survey of attacks and defense mechanism is given in [3, 35].

3. Proposed Attack

This section describes the proposed partial face tampering attack. The effect of the proposed attack on the performance of face recognition algorithms is evaluated with OpenFace [4] and VGG-Face [30] networks. Analysis is performed with respect to the deterioration in the performance of a face recognition system i.e., degradation in the verification accuracy of the system. Section 3.1 describes the partial face tampering attack, Section 3.2 presents the database and protocol, and Section 3.3 shows the effect of the proposed attack.

3.1. Partial Face Tampering Attack

Two different approaches are followed for generating tampered samples using partial face tampering attack. The first approach is referred as Replacement of Facial Regions (RFR) and the second approach as Morphing of Facial Regions (MFR). The details of the approaches are given below.

Replacement of Facial Regions:
In this approach, three different facial regions namely, eyes, mouth, and nose of an input image are replaced with the corresponding regions of another image (termed as source image) to generate the tampered samples. Each tampered sample contains one tampered region. Let $\mathbf{I}_i$ be the input image of subject i and $\mathbf{I}_j$ be the source image of subject j.

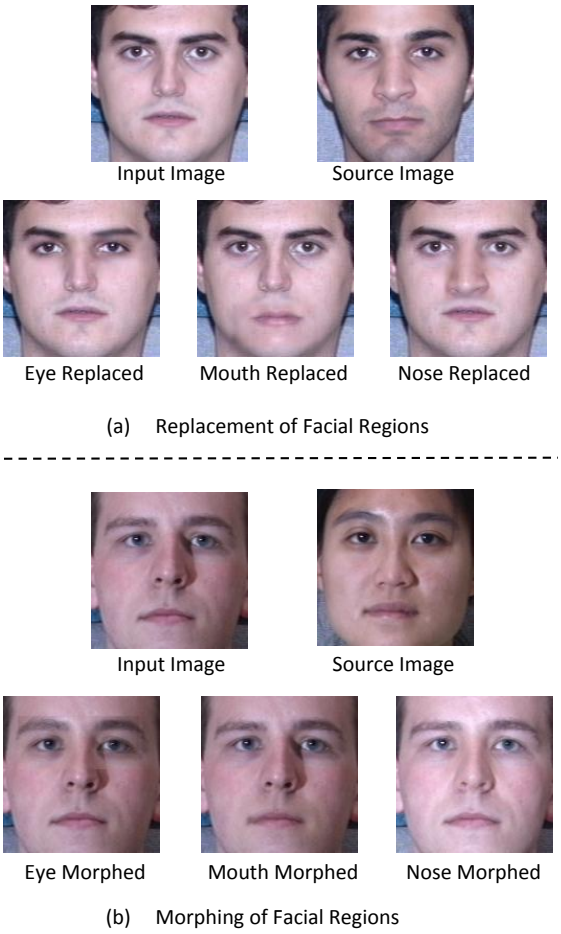

Figure 2. Sample images representing (a) Replacement of Facial Regions, (b) Morphing of Facial Regions.

RFR approach can be expressed as:

$$\mathbf{I}_{i,k} = \mathbf{I}_{j,k} \tag{1}$$

where, $\mathbf{I}_{i,k}$ is the k^{th} region of subject i and $\mathbf{I}_{j,k}$ is the k^{th} region of subject j. In order to replace the facial regions, Viola-Jones face detector [39] is used to locate eyes, mouth, and nose regions. Bounding box corresponding to the located regions are used to crop the facial regions from the source image and replaced with the input image. Further, edges of the replaced regions are smoothen out using Gaussian filtering. Figure 2(a) shows some samples generated using RFR approach. Three different categories of tampered images are created using the RFR approach: (i) eye full part, (ii) mouth full part, and (iii) nose full part, representing the replacement of eyes, mouth, and nose regions respectively.

Morphing of Facial Regions:

In this approach, eyes, mouth, and nose regions of an input image are morphed with the source image. For morphing of the facial regions, two different blending proportions, 0.4 and 0.5 are used. Blending proportion refers to the percentage of features of the source image blended with the input image. Similar to the RFR approach, let $\mathbf{I}_i$ be the input image of subject i and $\mathbf{I}_j$ be the source image of subject j. Morphing of Facial Regions (MFR) approach can be expressed as:

$$\mathbf{I}_{i,k} = \lambda \mathbf{I}_{i,k} + (1-\lambda)\mathbf{I}_{j,k} \tag{2}$$

where, $\mathbf{I}_{i,k}$ is the k^{th} region of subject i and $\mathbf{I}_{j,k}$ is the k^{th} region of subject j. λ is the parameter to control the blending proportion. Figure 2(b) shows some samples generated using MFR approach. Using this approach, six different categories of tampered images are created, namely, (i) eye morph 0.4, (ii) mouth morph 0.4, (iii) nose morph 0.4, (iv) eye morph 0.5, (v) mouth morph 0.5, and (vi) nose morph 0.5, representing morphing of eyes, mouth, and nose regions using 0.4 and 0.5 blending proportions respectively.

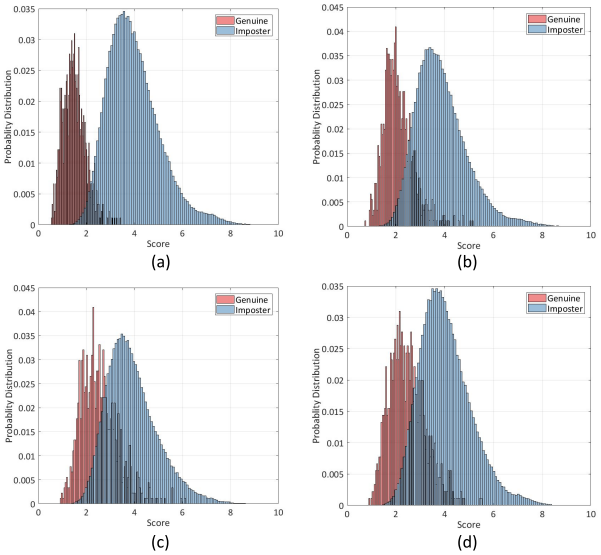

Figure 3. Genuine and imposter score distribution of OpenFace on Replacement of Facial Parts (RFR). (a) Score distribution of original probe images. (b-d) Score distribution of eyes, mouth, and nose replaced probe images respectively.

3.2. Database and Protocol

Experiments are performed on the CMU Multi-PIE [16] dataset. The dataset contains more than 75,000 images of 337 subjects. A subset of 226 subjects with 5 images per subject is used, out of which 4 are used to generate the tampered images and the remaining one image is used as the gallery image. The subset contains only frontal face images without glasses and proper illumination. As mentioned earlier, nine different categories of tampered images are generated using RFR and MFR approach. Each of the nine categories contain 904 (226×4) images.

Images are divided into gallery and 10 different probe sets. The gallery contains original images with a single im-

Table 1. Verification performance of OpenFace and VGG-Face in presence of visually similar tampered face images generated using RFR and MFR approach. The values indicate Genuine Accept Rate (%) at 1% False Accept Rate. MFR-0.4 represent results on images generated using 0.4 blending proportion and MFR-0.5 using 0.5 blending proportion.

Model		RFR			MFR-0.4			MFR-0.5		
	Original	Eye	Mouth	Nose	Eye	Mouth	Nose	Eye	Mouth	Nose
OpenFace	85.91	34.89	17.05	27.00	48.64	66.32	52.97	34.46	45.10	42.02
VGG-Face	99.97	53.59	94.97	66.90	92.02	99.77	97.53	70.91	99.60	90.98

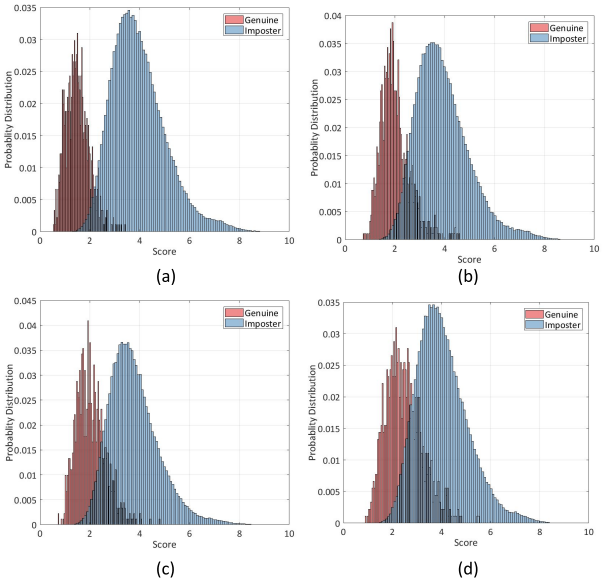

Figure 4. Genuine and imposter score distribution of OpenFace on Morphing of Facial Regions (MFR) with 0.5 blending proportion. (a) Score distribution of original probe images. (b-d) Score distribution of eyes, mouth, and nose morphed probe images respectively.

age per subject. Each probe set contains tampered images of a specific category resulting in nine different probe sets with an additional probe set containing the original counterpart of the tampered images. Each image in the probe set is matched with all the images in the gallery. The resulting score matrix of size 904×226 is used to determine the verification performance.

3.3. Effect of the Proposed Attack on Face Recognition

OpenFace and VGG-Face networks are utilized to determine the verification performance in the presence of tampered face images generated using RFR and MFR approaches. Features are extracted using the pre-trained models of the aforementioned deep networks, and Euclidean distance is computed between the probe and gallery images to generate the score matrix. Table 1 summarizes the effect of tampered face images on OpenFace and VGG-Face networks. As shown in Table 1, at 1% False Accept Rate (FAR), the Genuine Accept Rate (GAR) drops by approximately 51%, 68%, and 58% corresponding to the replacement of eyes, mouth, and nose regions respectively using

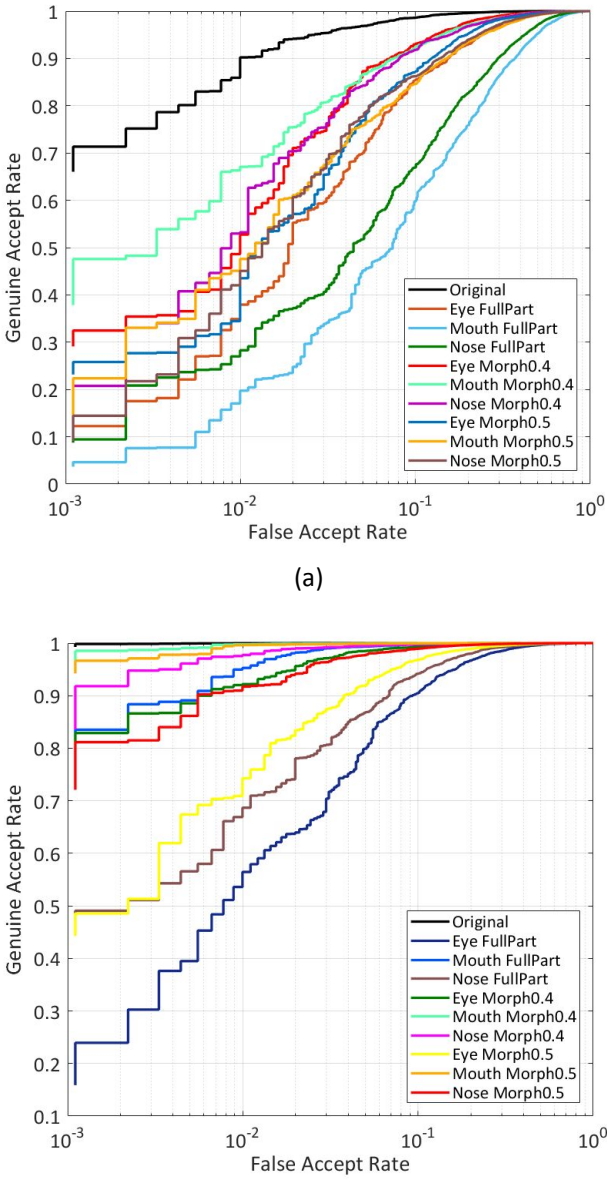

Figure 5. ROC plot of (a) OpenFace (b) VGG-Face, under the effect of Replacement of Facial Regions and Partial Morphing tampering artifacts.

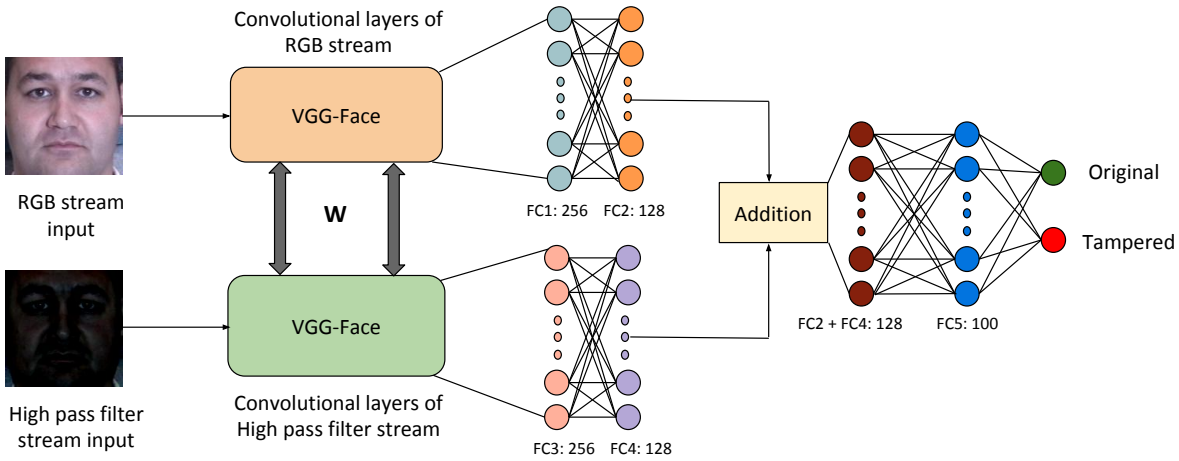

Figure 6. Proposed Partial Face Tampering Detection (PFTD) network. The RGB stream captures the inconsistencies like contrast difference and the high pass filter stream captures the local inconsistencies in the eyes, mouth, and nose regions.

OpenFace. VGG-Face shows a similar sharp drop in GAR corresponding to the replacement of facial parts. The genuine and imposter score distributions are shown in Figure 3. It is observed that the overlap increases by replacing the the facial regions. This, in turn, results in the sharp drop in GAR. It is asserted that since eyes, mouth, and nose are the most important and discriminative regions used by the face recognition algorithms [37, 42], therefore tampering these regions causes degradation in the performance.

Similar to the replacement of facial regions, morphing of facial regions also degrades the verification performance of the networks. The drop in verification performance increases with increasing blending proportion. For instance, the performance of OpenFace drops from 85.91% to 48.64%, 66.32%, and 52.97% corresponding to the morphing of eyes, mouth, and nose regions respectively using MFR-0.4 approach. The performance further drops to 34.46%, 45.10%, and 42.02% respectively using MFR-0.5 approach. However, the drop in GAR is not as significant as the replacement of facial regions. The reason being the presence of partial features of the genuine identity. The drop in verification performance indicates that minor changes in facial regions could mislead the existing systems and pose new challenges to the recognition systems. Figure 4 shows the genuine and imposter score distribution of OpenFace on MFR using 0.5 blending proportion. The increase in overlap between genuine and imposter score distribution emphasizes the degradation in the performance of the existing systems. Figure 5 shows the Receiver Operating Characteristic (ROC) Curve of OpenFace and VGG-Face under the effect of RFR and MFR tampering artifacts. The drop in GAR indicates that deep models are not robust to visually similar tampered face images generated using RFR and MFR approaches. It is, therefore, necessary to detect such attacks.

4. Detection of Partial Face Tampering Attack

The previous section shows that partial face tampering attack can degrade the performance of deep networks. This demands the necessity of a defense network for detecting such attacks to make the face recognition systems more robust towards tampering attacks. Therefore, a Partial Face Tampering Detection (PFTD) network is proposed for the detection of tampered samples. The performance of the proposed PFTD network is evaluated for detecting partial face tampering attack and compared with the existing deep models. Further, the robustness of PFTD network is evaluated for detecting unseen tampering attacks. Section 4.1 gives the details of the proposed PFTD network, Section 4.2 presents the implementation details, Section 4.3 discuss the experimental details and analysis, Section 4.4 shows the ablation study and Section 4.5 evaluates the robustness of the proposed PFTD network.

4.1. Proposed Partial Face Tampering Detection Network

The proposed PFTD network uses a combination of raw input and high pass filtered version of the input image for the detection of tampered images. The network has two streams namely, RGB stream and high pass filter stream. The RGB stream helps to capture the inconsistencies at the boundaries of the tampered regions or the contrast difference introduced in the image. On the other hand, the high pass filter stream captures the inconsistencies in the local regions such as eyes, mouth, and nose regions. The intuition behind using the high pass filter stream is that the artifacts introduced by the smoothing operations are better captured in the residual domain [32].

Figure 6 shows the proposed PFTD network. In the proposed network, VGG-Face is adopted by removing the top

layers within the two-stream network. As shown in Figure 6, weights are shared among the convolutional layers of the two streams. Two dense layers are added corresponding to each stream. The final layers of the two streams are added and followed by a common dense layer. During training, the first few layers of the VGG-Face network are frozen and the remaining layers along with the fully-connected layers are updated.

Let $\mathbf{X}$ be the training set with n number of images.

$$\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, ...\mathbf{X_n}\} \tag{3}$$

where, each $\mathbf{X_i}$ belongs to one of the two classes, namely, C_1 representing the $Original$ class and C_2 representing the $Tampered$ class. $Tampered$ class contains images created using RFR and MFR approaches. Let $\mathbf{X_i^{RGB}}$ be the input to the RGB stream and $\mathbf{X_i^{HPF}}$ be the input to the high pass filter stream. The output score of an input image X_i is represented as:

$$P(C_j|\mathbf{X_i}) = f(\mathbf{X_i^{RGB}}, \mathbf{X_i^{HPF}}, \mathbf{W}, b) \tag{4}$$

where, $P(C_j|\mathbf{X_i})$ is the probability of predicting image $\mathbf{X_i}$ to class C_j. $\mathbf{W}$ is the weight matrix and b is the bias. The network is trained with the following loss function:

$$L_{tot} = L_c + \delta L_m \tag{5}$$

where, L_c is the cross-entropy loss and L_m is the mean squared error. δ is a constant and set as 2 during the experiments. L_c is mathematically represented as:

$$L_c = -ylog(P) + (1-y)log(1-P) \tag{6}$$

where, y is the binary indicator if class label C_j is the correct classification for input image $\mathbf{X_i}$ and P is the probability of predicting $\mathbf{X_i}$ to C_j. L_m is mathematically represented as:

$$L_m = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i}) \tag{7}$$

where, y_i is the true class and $\hat{y_i}$ is the predicted class. Cross-entropy loss performs well for classification task and mean squared error give more penalty to the incorrect outputs due to the squared term. Therefore, a combination of the two loss functions is used to train the network.

4.2. Implementation Details

During training of the proposed PFTD network, the last five convolutional layers of the two streams followed by the dense layers are trained with RMSprop optimizer, and the learning rate is set to 0.00005. The network is trained for 50 epochs with a batch size of 128. ReLU activation function [41] is used in the dense layers. Further, experiments are performed on Tensorflow with Nvidia GTX 1080Ti GPU.

Table 2. Mean classification accuracy (%) of the existing and proposed models for the task of detecting partial face tampering attack.

	Models	**Classification Accuracy**
Existing	VGG16 [36]	54.06 ±0.01
	VGG-Face [30]	71.44 ±0.02
	OpenFace [4]	63.00 ±0.02
Fine-tuned	VGG16	70.33 ±0.05
	VGG-Face	81.78 ±0.06
	OpenFace	79.44 ±0.03
Proposed	RGB stream	87.61 ±0.02
	High pass filter stream	82.00 ±0.02
	PFTD	**91.44 ±0.01**

4.3. Experimental Details and Analysis

For experimental evaluation, five-fold cross-validation is performed with four folds for training and one fold for testing. The training set contains a total of 1440 images with 720 images belonging to the 'original' class and rest 720 to the 'tampered' class. 'Tampered' class contains an equal proportion of all nine variations of tampered images mentioned in section 3.1. For evaluating the performance of the existing deep networks, pre-trained models of VGG16 [36], VGG-Face [30], and OpenFace [4] are used. Features extracted using these pre-trained deep models are used to train a Support Vector Machine (SVM) [8]. First three rows of Table 2 shows the mean classification accuracy with the standard deviation of the five folds using the aforementioned deep models. From Table 2, it is observed that existing deep models do not perform well in detecting tampered images. Among the existing models, VGG-Face performs best. Further, existing models are fine-tuned on tampered samples generated using RFR and MFR approaches. It is observed from Table 2 that fine-tuning of the existing models enhances the performance. For instance, the classification accuracy increases by 16.27%, 10.34%, and 16.44% using fine-tuned VGG16, VGG-Face, and OpenFace models respectively. Fine-tuning helps the network to learn the tampering specific discriminative features to distinguish the tampered images from the original ones.

The performance of the proposed network is shown in the last row of Table 2. Further experiments are performed by ablating the high pass filter stream and ablating the RGB stream. Seventh and eighth rows of Table 2 shows the results for the same. It is observed that the proposed PFTD network improves the performance by 3.83% over the RGB stream and 9.44% over the high pass filter stream. As mentioned earlier, RGB stream helps to capture the contrast difference or the inconsistencies at the boundaries of the tampered regions. On the other hand, the high pass filter stream captures the local inconsistency in the eyes, mouth, and nose regions. It is asserted that training using the pro-

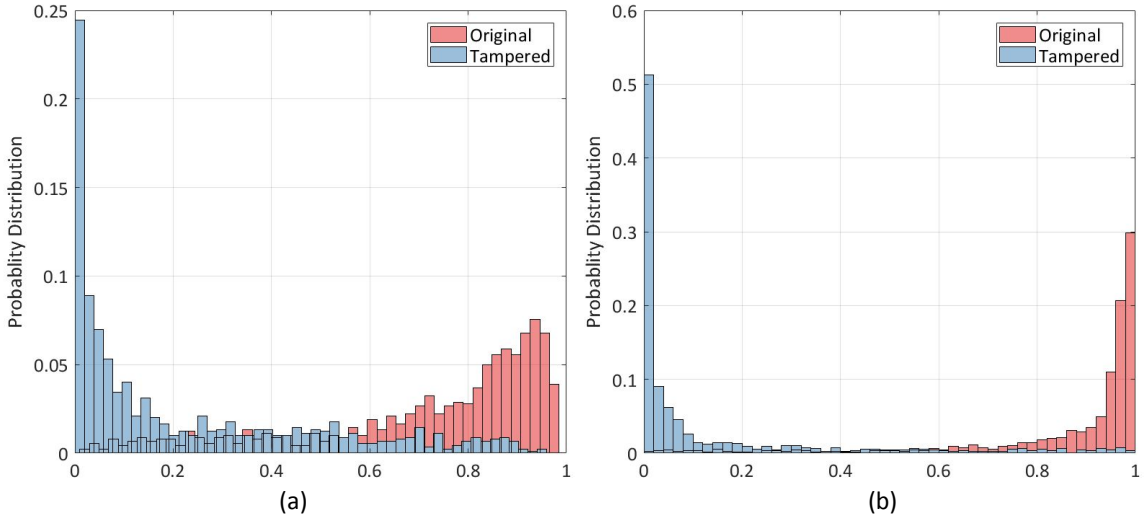

Figure 7. Score distribution of the original and tampered images using fine-tuned VGG-Face and proposed networks. (a) Fine-tuned VGG-Face and (b) Proposed.

Table 3. Confusion matrix (%) summarizing the results of the proposed model for the task of detecting adversarial face part tampering attack.

	Models		Ground Truth	
			Tampered	Original
Predicted	RGB stream	Tampered	82.78	7.56
		Original	17.22	92.44
	High-pass filter stream	Tampered	80.44	16.44
		Original	19.56	83.56
	PFTD	Tampered	**89.67**	6.78
		Original	10.33	**93.22**

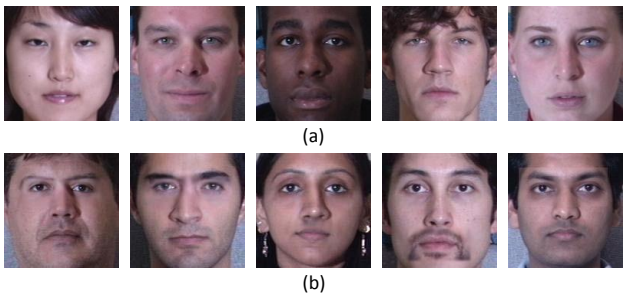

Figure 8. Sample images misclassified by the proposed Partial Face Tampering Detection network. (a) Original images classified as tampered. (b) Tampered images classified as original.

posed PFTD network captures the combined features which in turn help to further improve the performance of the network. Figure 7 shows the original and tampered score distribution of the fine-tuned VGG-Face and the proposed model. From Figure 7, it is observed that the proposed model reduces the overlap among the original and tampered classes and separates the two classes.

Confusion matrix of the RGB stream, high pass filter stream and proposed models is shown in Table 3. It is observed that the proposed network decreases both the False Reject Rate (FRR) and False Accept Rate (FAR). For instance, the proposed network decreases FRR by 6.89% and 9.23% while FAR by 0.78% and 9.66% from the RGB stream and high pass filter stream. Some misclassified images of both the classes are shown in Figure 8. The improved performance of the proposed network indicates its suitability towards the detection of the partial face tampering attack.

4.4. Ablation Study

To evaluate the effectiveness of the multi-loss function used to train the proposed PFTD network, two different ablation studies are performed. In the first experiment, the network is trained only with the cross-entropy loss and the performance of the network is evaluated. In the second experiment, the network is trained with mean squared error. Experiments with cross-entropy loss and mean squared error gives a classification accuracy of 90.61% and 89.88% respectively. In order words the classification accuracy degrades by 0.83% and 1.56% respectively as compared to the multi-loss function. This justifies the effectiveness of the combined loss function for the problem statement.

4.5. Robustness Analysis

In a real-world scenario, it is not pragmatic to assume knowledge about the type of tampering performed on an image. Therefore, the defense network must be robust towards unknown tampering attacks. In order to evaluate the per-

Table 4. Classification accuracy (%) of the proposed and fine-tuned VGG-Face models in detecting unknown tampering attacks (i.e., on DeepFake database [20]).

Models	Low Quality	High Quality
Fine-tuned VGG-Face	77.69	52.20
Proposed	**93.69**	**71.17**

formance of the proposed PFTD network, experiments are performed on the DeepFake database [20]. The database contains 640 tampered videos generated using Generative Adversarial Network (GAN) [13]. Among the 620 videos, 320 are of high quality and the rest 320 are of low quality. Experiments are performed on both types of videos using the PFTD network trained on the tampered samples generated using the RFR and MFR approaches. For experimental purpose, videos are converted to frames. The performance of the proposed PFTD network is compared with the best performing baseline fine-tuned model (i.e., VGG-Face) as shown in Table 2. Table 4 summarizes the result on unseen attack. From Table 4, it is observed that the proposed PFTD network performs equally well for low quality unseen attack videos. For high quality videos, the performance of both existing and PFTD is reduced, with PFTD performing better than fine-tuned VGG-Face. This indicates the robustness of the proposed network towards unknown tampering attacks.

5. Conclusion

Deep learning based face recognition systems are susceptible to digital attacks. In this research, partial face tampering attack is proposed, and the effect is evaluated on two state-of-the-art face recognition systems. The proposed attack replaces or morph facial regions of an input image with the source image. The images of same subjects are easily identified by humans. However, it is experimentally observed that existing deep face recognition systems are not able to identify the images of same subjects when the proposed partial face tampering attack is applied on the images. This in turn degrades the verification performance of the existing face recognition algorithms. The answer to the question asked in Figure 1 is given in Figure 9. As shown in the Figure 9, red rectangular boxes indicates the tampered regions.

Further, a Partial Face Tampering Detection network is proposed for the task of detecting the proposed attack, and the performance is compared with the baseline algorithms. The proposed network uses a combination of RGB input and a high pass filtered version of the input image to capture the inconsistencies among the original and tampered images. The proposed network enhances the detection performance by 20% and 9.66% from the best performing pre-trained and fine-tuned model respectively. In the future, the aim is to detect the tampered regions to develop robust algorithms for mitigation.

Figure 9. Images marked with red rectangular box are the tampered images with the replaced region inside the box.

6. Acknowledgements

A. Agarwal is partly supported by Visvesvaraya PhD Fellowship, and M. Vatsa and R. Singh are partly supported from the Infosys Center for AI at IIIT-Delhi. M. Vatsa is also partially supported by the Department of Science and Technology, Government of India through Swarnajayanti Fellowship.

References

[1] A. Agarwal, R. Singh, M. Vatsa, and A. Noore. Swapped! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE/IAPR International Joint Conference on Biometrics*, 2017. 1, 2

[2] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect. *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2018. 2

[3] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 2

[4] B. Amos, B. Ludwiczuk, J. Harkes, P. Pillai, K. Elgazzar, and M. Satyanarayanan. Openface: Face recognition with deep neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2016. 1, 2, 6

[5] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer. Detecting facial retouching using supervised deep learning. *IEEE Transactions on Information Forensics and Security*, 11(9):1903–1913, 2016. 1

[6] A. Bharati, M. Vatsa, R. Singh, K. W. Bowyer, and X. Tong. Demography-based facial retouching detection using subclass supervised sparse autoencoder. In *IEEE International Joint Conference on Biometrics*, pages 474–482, 2017. 1

[7] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 1, 2

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 6

[9] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *IEEE/IAPR International Joint Conference on Biometrics*, 2014. 2

[10] M. Ferrara, A. Franco, and D. Maltoni. On the effects of image alterations on face recognition accuracy. In *Face recognition across the imaging spectrum*, pages 195–222. 2016. 1

[11] M. Ferrara, A. Franco, and D. Maltoni. Face demorphing. *IEEE Transactions on Information Forensics and Security*, 13(4):1008–1017, 2018. 2

[12] A. Goel, A. Singh, A. Agarwal, M. Vatsa, and R. Singh. Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition. *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2018. 2

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 8

[14] G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *International Journal of Computer Vision*, 2019. doi: 10.1007/s11263-019-01160-w. 1

[15] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. *Association for the Advancement of Artificial Intelligence*, 2018. 1, 2

[16] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 3

[17] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017. 2

[18] M. Hildebrandt, T. Neubert, A. Makrushin, and J. Dittmann. Benchmarking face morphing forgery detection: application of stirtrace for impact simulation of different processing steps. In *International Workshop on Biometrics and Forensics*, 2017. 1

[19] H. Hosseini, Y. Chen, S. Kannan, B. Zhang, and R. Poovendran. Blocking transferability of adversarial examples in black-box learning systems. *arXiv preprint arXiv:1703.04318*, 2017. 2

[20] P. Korshunov and S. Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 8

[21] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *IEEE International Conference on Computer Vision*, 2017. 1

[22] C. Kraetzer, A. Makrushin, T. Neubert, M. Hildebrandt, and J. Dittmann. Modeling attacks on photo-id documents and applying media forensics for the detection of facial morphing. In *ACM Workshop on Information Hiding and Multimedia Security*, 2017. 1

[23] X. Li and F. Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *IEEE International Conference on Computer Vision*, 2017. 2

[24] Y. Li, M.-C. Chang, and S. Lyu. In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security*, 2018. 2

[25] J. Lu, T. Issaranon, and D. A. Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *IEEE International Conference on Computer Vision*, 2017. 2

[26] 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. 1

[27] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017. 2

[28] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1773, 2017. 2

[29] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1, 2

[30] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015. 1, 2, 6

[31] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch. Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 2

[32] Y. Rao and J. Ni. A deep learning approach to detection of splicing and copy-move forgeries in images. In *IEEE International Workshop on Information Forensics and Security*, 2016. 5

[33] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 1

[34] U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch. On the vulnerability of face recognition systems towards morphed face attacks. In *IEEE International Workshop on Biometrics and Forensics*, 2017. 1, 2

[35] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7:23012–23026, 2019. 2

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[37] 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. 5

[38] J. Su, D. V. Vargas, and S. Kouichi. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*, 2017. 1, 2

[39] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I511–I518, 2001. 3

[40] 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. 1

[41] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 6

[42] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003. 5