Fake Face Detection Methods: Can They Be Generalized?

1st Ali Khodabakhsh, 2nd Raghavendra Ramachandra, 3rd Kiran Raja, 4th Pankaj Wasnik, 5th Christoph Busch
Department of Information Security and Communication Technology,
Norwegian University of Science and Technology
Gjovik, Norway
{ali.khodabakhsh, raghavendra.ramachandra, kiran.raja, pankaj.wasnik, christoph.busch}@ntnu.no

***Abstract*—With advancements in technology, it is now possible to create representations of human faces in a seamless manner for fake media, leveraging the large-scale availability of videos. These fake faces can be used to conduct personation attacks on the targeted subjects. Availability of open source software and a variety of commercial applications provides an opportunity to generate fake videos of a particular target subject in a number of ways. In this article, we evaluate the generalizability of the fake face detection methods through a series of studies to benchmark the detection accuracy. To this extent, we have collected a new database of more than $53,000$ images, from 150 videos, originating from multiple sources of digitally generated fakes including Computer Graphics Image (CGI) generation and many tampering based approaches. In addition, we have also included images (with more than $3,200$) from the predominantly used Swap-Face application that is commonly available on smartphones. Extensive experiments are carried out using both texture-based handcrafted detection methods and deep learning based detection methods to find the suitability of detection methods. Through the set of evaluation, we attempt to answer if the current fake face detection methods can be generalizable.**

***Index Terms*—Fake Face, Presentation Attack Detection, Dataset, Generalization, Transfer Learning**

I. Introduction

Face biometrics are widely deployed in various applications as it ensures reliable and convenient verification of a data subject. The dominant application of face recognition is for logical or physical access control to for instance restricted security areas. Implicitly the human visual system applies face recognition to determine, which data subject is the communication partner be it in a face to face conversation or be it in consuming messages while observing a media stream (e.g. news channel). With recent advances in deep learning, it is now possible to seamlessly generate manipulated images/videos in real-time using technologies like image morphing, Snap-Chat, Computer Generated Face Image (CGFI), Generative Adversarial Networks (GAN) and Face2Face [1]. These technologies enable an attacker to manipulate the face image either by swapping it with another face or by pixel-wise manipulation to generate a new face image/video. It is well demonstrated in the literature that face recognition techniques fail drastically in detecting generated fake faces [2]. Further fake face samples can also be shared by intention with the social media, in order to spread the fake news associated with the target subject. The challenge is not only posed to the biometric systems but also to the general media perception on social media. Thus it is of paramount importance to detect faked face representations to reduce the vulnerability of biometrics systems and to reduce the impact of manipulated social media content.

(a) Bona fide (b) Face retargeting[1]

(c) DeepFake[2] (d) CGI[3]

Fig. 1: Examples of different fake faces in contrast to the bona fide presentation.

Traditional biometric systems have addressed this problem of detecting the fake faces using Presentation Attack Detection (PAD) schemes [3], [4]. PAD schemes in the earlier works have investigated and provided remedial measures focused on both attacks with low-cost artefacts (e.g. print, display, and wrap) and high-cost artefacts (like silicon masks). Another kind of attacks based on face morphing takes face images of two different data subjects to generate a new morphed face

[1]Pinscreen: http://www.pinscreen.com/
[2]https://www.fakeapp.org/
[3]"We the people": http://www.macinnesscott.com/vr-art-x

image which can practically match both the subjects [2]. Yet another and recently created method of generating a faked face image/video was presented in [1] that can be used to introduce a personation attack on the target subject. The personation attack can be constructed by the re-enactment process, transferring the facial expressions from the source actor to a target actor, resulting in the manipulated images/video. This generated facial sample through such procedures is referred to as the fake face [5], [6]. The generated content shows high sample quality of images/videos, which is difficult to detect even for trained forensic examiners [6]. There are recent additions to generate fake face images that include the use of GAN, CGI, Face2face, and others which are highly realistic. The reliable detection of such fake face images is challenging due to the process of re-enactment. This results in infinitesimal variation in the face images that challenges the conventional forensics methods based on extracting edge discontinuities and texture information in spotting manipulated images.

To the best of our knowledge, there exists only one work that has attempted to detect fake faces, which were using only one type of fakes, generated by Face2Face application [6]. In their work [6], pre-trained deep Convolutional Neural Network (CNN) based approaches are evaluated on the newly constructed fake face image database. The results reported in [6] show good detection performance of the pre-trained Xception CNN that can be attributed to the fact that both fake face generation and detection are carried out on the training and testing subset of one particular dataset (FaceForensics). While this is an important first step, we need to anticipate that with the evolution of computer vision technologies, fake faces can also be generated using alternative and newer methods. Thus, it is necessary to provide an insight into the generalization of the methods that are used to detect the fake faces to measure the reliability.

In this work, we present a comprehensive and exploratory study on the generalizability of different fake face detection methods based on both recent deep learning methods and conventional texture descriptor based methods. To this extent of studying generalizability, we present a new database created using diverse methodologies for generating fake faces. Further, we also propose the protocols to effectively evaluate the generalizability of both texture based and deep learning based methods. The main contributions of this paper in fake face detection are:

- A new database which we hereafter refer as *Fake Face in the Wild (FFW)* database with more than $53,000$ images (from 150 videos) assembled from public sources (YouTube) is introduced. This database shows the largest diversity of different fake face generation methods provided so far.
- In the view of limited public databases available for this key research area, the newly created database will be made available for the public along with the publication of this paper.
- Comprehensive evaluation of 6 different algorithms that include various kinds of deep learning methods such as AlexNet [7], VGG19 [8], ResNet50 [9], Xception [10], GoogLeNet/Inceptionv3 [11], and texture based methods based on Local Binary Patterns (LBP) with Support Vector Machine (SVM).
- Extensive experiments providing insights on the generalization of the algorithms for unseen fake faces are presented. Specifically, fake faces generated using three different methods such as CGI, FakeApp, face swap, etc are considered.

Category	Type	# of videos
CGI	Full	50
	Head	22
Tampering	Face (FakeApp)	50
	Face (Other)	28
Total		150

TABLE I: Fake Face in the Wild Dataset (FFW) broad statistics. CGI faces were generated using several different graphics engines. Face (FakeApp) were generated in multiple resolutions and with different settings. Face (Other) category includes Face replacement, part of face splicing, and partial CGI faces, some of which were done manually, others automatically (see Figure 3 for examples).

II. Fake Face in the Wild Dataset (FFW)

This section presents the details of the newly constructed database. To simulate the performance of fake face detection methods in the wild, a set of videos from a public video sharing website (YouTube) is collected. This dataset is collected with the special focus on digitally created contents, generated with recently developed technologies. These videos include a wide array of fake images generated through CGI, GANs, manual and automatic image tampering techniques, and their combinations, due to the widespread use of these methodologies. CGI is considered in this work due to the wide availability and the ease of creation of high-quality fake face images that include images of variable sizes. *The key motivation in creating this database can be attributed to non-available public databases for either devising detection methods or the study of generalizability.* This work, therefore, facilitates further research by making the dataset publicly available along with the paper.[4].

Table I shows a summary of the videos in the FFW dataset. The dataset is created using videos of variable duration ranging from 2 seconds that corresponds to 60 frames up-to 74 seconds that corresponds to more than $2,000$ frames. The videos are carefully selected to have a resolution of at least 480p and above and are manually checked for assuring the quality to avoid images with visible artifacts, face poses, degraded illumination on faces and resolution. The constructed dataset consists of 150 videos, of which 85 videos broadly pertain to face images manipulated via image tampering (e.g., splicing, replacing, etc) and 65 corresponds to the use of CGI. The database thus consists of $53,000$ images. In order to have bona

[4]Download information available at http://ali.khodabakhsh.org/ffw/

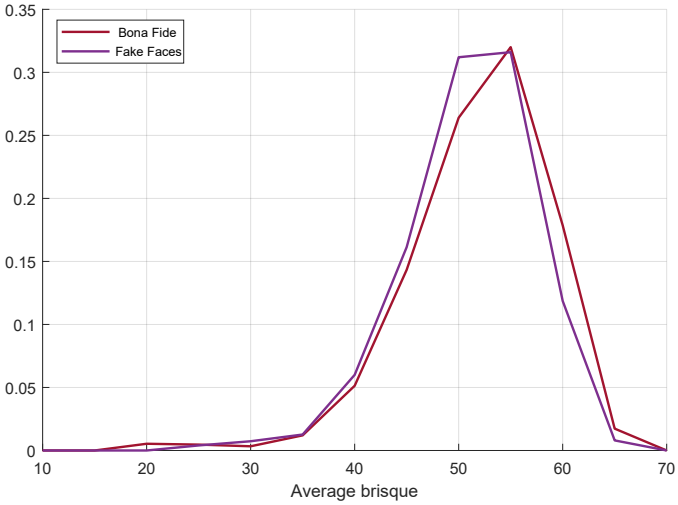

Fig. 2: Distribution of BRISQUE quality scores for the Fake Faces in the Wild (FFW) dataset.

fide samples for the evaluation, we have employed publicly available face forensic database [6] resulting in a total of $78,500$ bona fide samples from 150 videos.

To evaluate the performance on the newly created database, the quality measures are taken into consideration by processing the database through the same compression algorithm such that the quality of both fake and bona fide samples are consistent. This further avoids misleading detection error rates that for instance can be attributed to compression artefacts and bias the detection methods. Figure 2 shows the distribution of the average BRISQUE quality assessment [12] measured for FFW database indicating high overlap of the distribution justifying the similar quality. A sample set of images from the FFW dataset can also be seen in Figure 3.

III. Fake Face Detection Techniques

With the goal of detection of a wide range of forged/CG/tampered audiovisual content, many methods originating from image forensics and biometrics presentation attack detection can be adapted. In this perspective, widely used texture-based method - Local Binary Patterns (LBP) and a set of CNN based systems are considered. The selection of CNN architectures AlexNet [7], VGG19 [8], ResNet50 [9], Xception [10], and GoogLeNet/Inceptionv3 [11] is based on the recent works demonstrating very high performance for various tasks. The parameters are optimized when possible on the training data and the details of parameter tuning is presented in IV-B.

IV. Experimental Evaluation

This section presents the experimental evaluation of the FFW dataset. The experiment protocols are designed in accordance with protocols advised in [6]. We present the evaluation of detecting known attacks followed by detecting unknown attacks.

A. Evaluation Metrics

We present the detection error rates in terms of Equal Error Rates (EER) to provide performance in the lines of earlier work. We further supplement the results using the ISO/IEC 30107-3 [13] with Attack Presentation Classification Error Rate (APCER) and Bona fide Presentation Classification Error Rate (BPCER) as described in [13].

B. Experimental Protocol

To effectively evaluate the fake detection methods, we divide the whole database to have three different disjoint partitions such as training set, development set, and testing set. The training set is adopted from the FaceForensics database [6] that has $7,040$ bona fide and $7,040$ fake face samples. The training set is used to fine tune the pre-trained deep CNN networks. To effectively fine-tune the networks and avoid overfitting, we employ 5 different types of data augmentation on each of the training images that includes translation and reflection. The learning rates of the last layer are boosted such that weights of the earlier layer are not affected and the weights of the last layer are adapted for the new training data. Thus, we have used the *weight learning rate factor* as 10 and *bias learning rate factor* as 20. For the texture based Local Binary Patterns (LBP) [14], the histogram is extracted using (8,1) neighborhoods with a block size of 40 pixels. The training dataset is used to train the SVM classifier.

The development dataset is comprised of $1,500$ bona fide and $1,500$ fake face samples that are taken from the validation set of FaceForensics database [6]. This dataset is used to fix the operating thresholds such as Equal Error Rates (EER). The testing dataset consists of three specific kinds: (1) *To evaluate known artefacts - TestSet-I* - Test set corresponds to test set of FaceForensics database [6] that comprised of $1,500$ bona fide and $1,500$ fake face samples. This dataset is particularly used to understand the detection performances of known attacks. (2) *To evaluate unknown artefacts - TestSet-II* - The test set in this case consists of a newly constructed FFW dataset. In order to be inline with known attacks, this set is comprised of $1,500$ bona fide and $1,500$ fake face samples. (3) *To evaluate unknown artefacts - TestSet-III* - This test set comprises of $1,776$ bona fide samples and $1,576$ fake faces generated using FaceSwap and SwapMe application proposed by [15].

While *TestSet-I* focuses on measuring the performance of the detection algorithms, *TestSet-II* and *TestSet-III* are used to measure the generalizability of the detection techniques. It has to be noted that none of these sets (*TestSet-II* and *TestSet-III*) are used either for training, fine-tuning or validation process.

V. Results and Discussion

The detailed results and the obtained performance are provided in this section.

A. Performance on the Known Fake Face Attacks (TestSet-I)

The performance of texture- and CNN-based methods on known attacks (TestSet-I) are summarized in Table II and Table III. Following are the main observations:

Fig. 3: Examples from Fake Faces in the Wild (FFW) dataset. Top row: CGI full scene. Middle row: Deepfakes. Bottom row from left to right: Head CGI x2, Face replacement x2, Face CGI x2, Part of face splicing x2.

		Accuracy ± CI
Texture-based	LBP	96.33% ± 0.69%
CNN-based	AlexNet	95.83% ± 0.73%
	VGG19	98.30% ± 0.47%
	ResNet	98.43% ± 0.45%
	Xception	98.70% ± 0.41%
	Inception	99.60% ± 0.23%

TABLE II: The accuracy of texture- and CNN-based classifiers on the TestSet I dataset along with their confidence interval (CI).

	APCER	BPCER	EER
LBP	3.80% ± 0.99%	2.87% ± 0.86%	3.33%
AlexNet	7.80% ± 1.38%	1.73% ± 0.67%	3.73%
VGG19	2.47% ± 0.80%	0.47% ± 0.35%	1.40%
ResNet	2.27% ± 0.77%	0.47% ± 0.35%	1.40%
Xception	2.47% ± 0.80%	0.13% ± 0.19%	1.07%
Inception	0.67% ± 0.42%	0.47% ± 0.35%	0.53%

TABLE III: Performance of the systems on known fake faces from TestSet I. The threshold is computed on the development database.

- CNN-based methods perform well and except for AlexNet, provide a detection accuracy of over 98%. In contrast, LBP features classified with SVM have the accuracy of 96% on the test data.
- In the benchmark of the CNN networks, the Inception network gives the best performance by a large margin.
- The low error rates in accord with a low EER error confirm the stability of the selected threshold point for decision. However, deviation from the selected operating point towards lower BPCER and higher APCER is visible in the results, suggesting slight inaccuracy in EER threshold estimation.

	APCER	BPCER	EER
LBP	89.00% ± 1.62%	2.87% ± 0.86%	48.73%
AlexNet	91.47% ± 1.44%	1.73% ± 0.67%	32.13%
VGG19	90.73% ± 1.50%	0.47% ± 0.35%	29.40%
ResNet	89.53% ± 1.58%	0.47% ± 0.35%	30.33%
Xception	93.20% ± 1.30%	0.13% ± 0.19%	26.87%
Inception	91.93% ± 1.41%	0.47% ± 0.35%	27.47%

TABLE IV: Performance of the systems on unknown attacks from TestSet II. The threshold is computed on the development database.

B. Performance on the Unknown Fake Face Presentations (TestSet-II)

Following the good performance of all neural network solutions along with the LBP features, the generalizability of the learned classifiers are examined on the collected dataset of matching size as shown in Table IV and the observations are:

- The performance of all systems in terms of APCER errors drops significantly, rendering the systems ineffective, classifying most images as bona fide.
- A closer look at the EER values for these systems shows much better than random performance of CNN-based models on the Unknown dataset.
- It can be concluded that the performance of the CNN-based systems is very poor because of the low performance at the selected operating point.

To illustrate this further, the score histogram of the known and unknown attacks are presented in Figures 4 and 5 for LBP-SVM and Inception networks respectively. The dotted vertical line indicates the threshold computed on the development database that corresponds to the EER. Figure 4 shows the inability of the system in distinguishing unknown attacks by a significant overlap between the bona fide distribution and the distribution of scores from the unknown attacks. However, a

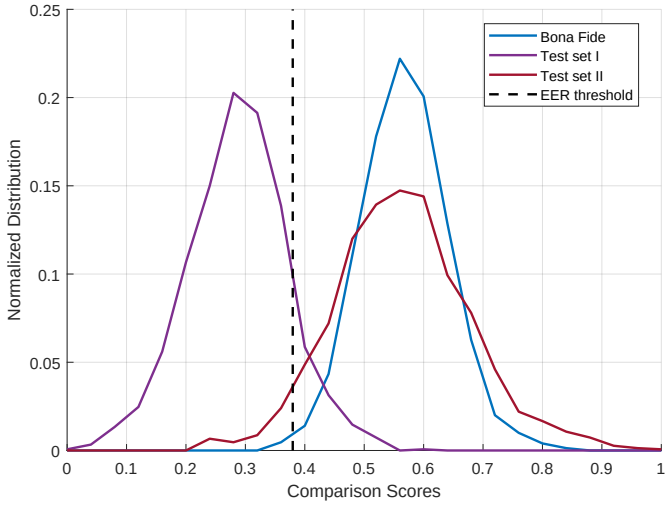

Fig. 4: LBP-SVM system comparison score distribution on TestSets I and II.

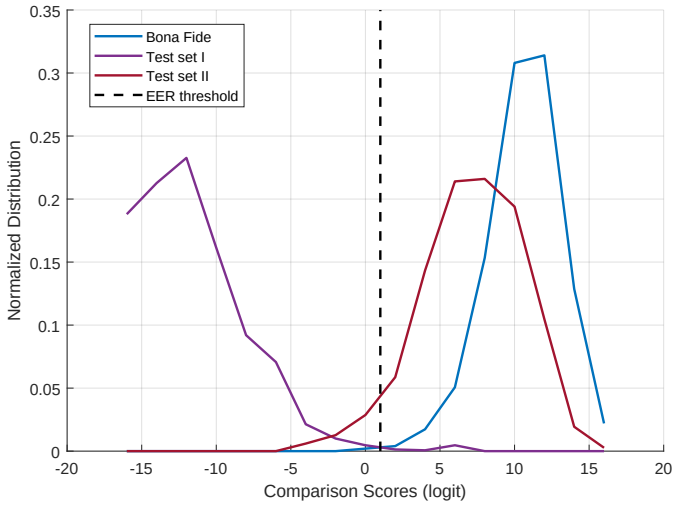

Fig. 5: Inceptionv3 system comparison score distribution on TestSets I and II.

close look into Figure 5 shows that even though the network is capable of discriminating between unknown attacks and the bona fide to some extent, the weak placement of the decision boundary causes the network to fail. *By setting the threshold of the system to the EER point on the known attacks, even though the system shows optimal performance for the known attacks, it also becomes vulnerable to new types of attacks, where the separability may be less.*

1) Performance on each Sub-Type of Attacks: To have a closer look at the capability of CNNs in generalization, EERs for each type is calculated separately and reported in Table V.

- From these results, it is visible that the networks perform better in detecting CGI compared to contents generated by FakeApp, or other techniques.
- These results indicate that even though the networks were not trained to detect CGI specifically, they are still somewhat effective for detecting of CGI videos.

	Full CGI	Image Manipulation	
		FakeApp	Other
AlexNet	32.60%	28.80%	34.37%
VGG19	28.00%	31.20%	28.60%
ResNet	28.80%	28.37%	34.40%
Xception	23.60%	25.20%	31.20%
Inception	23.40%	27.40%	31.40%

TABLE V: CNN performances in terms of EER on subcategories, corresponding to Table I.

	APCER	BPCER	EER
LBP	90.16% ± 1.50%	3.43% ± 0.86%	46.06%
AlexNet	94.04% ± 1.19%	5.01% ± 1.04%	43.02%
VGG19	97.27% ± 0.82%	2.31% ± 0.71%	44.93%
ResNet	89.40% ± 1.55%	8.22% ± 1.30%	43.79%
Xception	93.15% ± 1.27%	3.43% ± 0.86%	40.99%
Inception	71.64% ± 2.27%	22.58% ± 1.98%	46.39%

TABLE VI: Performance of the systems on FaceSwap/SwapMe dataset from TestSet III. The threshold is computed on the development database.

C. Performance on the FaceSwap/SwapMe Dataset (TestSet-III)

To investigate the transferability of the generalization ability of the networks on the unknown data of a widely different type, experiments were done on a filtered subset of the FaceSwap/SwapMe dataset as shown in Table VI.

- The APCER and EER scores present a further drop in performance.
- These results indicate the lack of transferability of the learned classifiers to the general face forgery classification cases.

VI. Conclusion and Future Work

The advancement of image manipulation and image generation techniques have now provided the ability to create seamless and convincing fake face images. The challenging nature of data both for visual perception and algorithmic detection is provided in recent works. The key problem that was not considered up until now is the evaluation of generalizability on existing fake face detection techniques. In order to answer the question of generalizability, in this work, we have created a new database which we refer to as Fake Face in the Wild (FFW) dataset containing $53,000$ images from 150 videos that are publicly available. The key observation from this work throws light on deficiencies of detection algorithms when unknown data is presented. This observation holds for both texture descriptors and deep-learning methods, which yet cannot meet the challenge of detecting fake faces. This analysis further emphasizes the importance of validation of detectors across multiple datasets. Proposed detectors that lack such validation can show misleadingly high performances while having limited applicability, and provide little contribution to the ongoing research. As such, advancements in fake face

detection technology call for the incorporation of proper cross-dataset validation in all future research as a requirement for publication.

The future work in the direction of fake face detection will involve the development of systematical methods for answering the generalization problem, and employment of multi-modal cues from fake face data.

REFERENCES

[1] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 2387–2395.

[2] R. Raghavendra, K. B. Raja, and C. Busch, "Detecting morphed face images," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sept 2016, pp. 1–7.

[3] S. Bhattacharjee and S. Marcel, "What you can't see can help you - extended-range imaging for 3d-mask presentation attack detection," in *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sept 2017, pp. 1–7.

[4] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Comput. Surv.*, vol. 50, no. 1, pp. 8:1–8:37, Mar. 2017. [Online]. Available: http://doi.acm.org/10.1145/3038924

[5] A. Khodabakhsh, R. Ramachandra, and C. Busch, "A taxonomy of audiovisual fake multimedia content creation technology," in *Proceedings of the 1st IEEE International Workshop on Fake MultiMedia (FakeMM'18)*, 2018, pp. –.

[6] 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.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *CoRR*, vol. abs/1610.02357, 2016. [Online]. Available: http://arxiv.org/abs/1610.02357

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015. [Online]. Available: http://arxiv.org/abs/1512.00567

[12] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.

[13] ISO/IEC 30107-3:2017, "Information technology - Biometric presentation attack detection - Part 3: Testing and reporting," International Organization for Standardization, Standard, Sep. 2017.

[14] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2002.1017623

[15] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE*, 2017, pp. 1831–1839.