

Celeb-DF (v2): A New Dataset for DeepFake Forensics

Yuezun Li¹, Xin Yang¹, Pu Sun², Honggang Qi² and Siwei Lyu¹

¹ University at Albany, State University of New York, USA

² University of Chinese Academy of Sciences, China

Abstract

AI-synthesized face-swapping videos, commonly known as DeepFakes, is an emerging problem threatening the trustworthiness of online information. The need to develop and evaluate DeepFake detection algorithms calls for large-scale datasets. However, current DeepFake datasets suffer from low visual quality and do not resemble DeepFake videos circulated on the Internet. We present a new large-scale challenging DeepFake video dataset, Celeb-DF¹, which contains 5,639 high-quality DeepFake videos of celebrities generated using improved synthesis process. We conduct a comprehensive evaluation of DeepFake detection methods and datasets to demonstrate the escalated level of challenges posed by Celeb-DF.

1. Introduction

A recent twist to the disconcerting problem of online disinformation is falsified videos created by AI technologies, in particular, deep neural networks (DNNs). Although fabrication and manipulation of digital images and videos are not new [11], the use of DNNs has made the process to create convincing fake videos increasingly easier and faster.

One particular type of DNN-based fake videos, commonly known as *DeepFakes*, has recently drawn much attention. In a DeepFake video, the faces of a *target* individual are replaced by the faces of a *donor* individual synthesized by DNN models, retaining the target’s facial expressions and head poses. Since faces are intrinsically associated with identity, well-crafted DeepFakes can create illusions of a person’s presence and activities that do not occur in reality, which can lead to serious political, social, financial, and legal consequences [7].

With the escalated concerns over the DeepFakes, there is a surge of interest in developing DeepFakes detection methods recently [4, 12, 16, 32, 20, 17, 26, 25, 22, 21, 23], with an upcoming dedicated global *DeepFake Detection Chal-*

*lenge*². The availability of large-scale datasets of DeepFake videos is an enabling factor in the development of DeepFake detection method. To date, we have the UADFV dataset [32], the DeepFake-TIMIT dataset (DF-TIMIT) [15], the FaceForensics++ dataset (FF-DF) [25]³, the Google DeepFake detection dataset (DFD) [10], and the FaceBook DeepFake detection challenge (DFDC) dataset [9].

However, a closer look at the DeepFake videos in existing datasets reveals some stark contrasts in visual quality to the actual DeepFake videos circulated on the Internet, which includes low-quality synthesized faces, visible splicing boundaries, color mismatch, visible parts of the original face, and inconsistent synthesized face orientations. These artifacts are likely the result of imperfect steps of the synthesis method and the lack of curating of the synthesized videos before included in the datasets. Moreover, DeepFake videos with such low visual qualities can hardly be convincing, and are unlikely to have real impact. Correspondingly, high detection performance on these dataset may not bear strong relevance when the detection methods are deployed *in the wild*.

In this work, we present a new large-scale and challenging DeepFake video dataset, *Celeb-DF*, for the development and evaluation of DeepFake detection algorithms. There are in total 5,639 DeepFake videos, corresponding more than 2 million frames, in the Celeb-DF dataset. The real source videos are based on publicly available YouTube video clips of 59 celebrities of diverse genders, ages, and ethnic groups. The DeepFake videos are generated using an improved DeepFake synthesis method. As a result, the overall visual quality of the synthesized DeepFake videos in Celeb-DF is greatly improved when compared to existing datasets, with significantly fewer notable visual artifacts, see Fig.1, and example videos in the supplementary materials. Based on the Celeb-DF dataset and other existing datasets, we conduct an evaluation of a large set of current DeepFake detection methods. This is the most comprehensive performance evaluation of DeepFake detection meth-

¹<https://github.com/danmohaha/celeb-deepfakeforensics>.

²<https://deepfakedetectionchallenge.ai>.

³The FaceForensics++ dataset contains other types of fake face videos. For the relevance of this work, we only consider the set of DeepFakes.



Figure 1. Example frames from the Celeb-DF dataset. Left column is the frame of real videos and right five columns are corresponding DeepFake frames generated using different donor subject.

ods to date. The results show that Celeb-DF is challenging to most of the existing detection methods, even though many of these DeepFake detection methods are shown to achieve high, sometimes near perfect, accuracy on previous datasets.

2. Backgrounds

2.1. DeepFake Detection Methods

Since DeepFakes become a global phenomenon, there has been an increasing interest in DeepFake detection methods. Most of the current DeepFake detection methods use data-driven deep neural networks (DNNs) as backbone.

Since synthesized faces are spliced into the original video frames, state-of-the-art DNN splicing detection methods, e.g., [33, 34, 18, 6], can be applied. There have also been algorithms dedicated to the detection of Deep-

Fake videos that fall into three categories. Methods in the first category are based on inconsistencies exhibited in the **physical/physiological** aspects in the DeepFake videos. The method in work of [16] exploits the observation that many DeepFake videos lack reasonable eye blinking due to the use of online portraits as training data, which usually do not have closed eyes for aesthetic reasons. Incoherent head poses in DeepFake videos are utilized in [32] to expose DeepFake videos. In [5], the idiosyncratic behavioral patterns of a particular individual are captured by the time series of facial landmarks extracted from real videos are used to spot DeepFake videos. The second category of DeepFake detection algorithms (e.g., [20, 17]) use **signal-level** artifacts introduced during the synthesis process such as those described in the Introduction. The third category of DeepFake detection methods (e.g., [4, 12, 22, 23]) are **data-driven**, which directly employ various types of DNNs

Dataset	# Real		# DeepFake		Release Date
	Video	Frame	Video	Frame	
UADFV	49	17.3k	49	17.3k	2018.11
DF-TIMIT-LQ	320*	34.0k	320	34.0k	2018.12
DF-TIMIT-HQ			320	34.0k	
FF-DF	1,000	509.9k	1,000	509.9k	2019.01
DFD	363	315.4k	3,068	2,242.7k	2019.09
DFDC	1,131	488.4k	4,113	1,783.3k	2019.10
Celeb-DF	590	225.4k	5,639	2,116.8k	2019.11

Table 1. Basic information of various DeepFake video datasets. *: the original videos in DF-TIMIT are from Vid-TIMIT dataset.

trained on real and DeepFake videos, not relying on any specific artifact.

2.2. Existing DeepFake Video Datasets

DeepFake detection methods require training data and need to be evaluated. As such, there is an increasing need for large-scale DeepFake video datasets. Table 1 lists the current DeepFake datasets.

UADFV: The UADFV dataset [32] contains 49 real YouTube and 49 DeepFake videos. The DeepFake videos are generated using the DNN model with FakeAPP [3].

DF-TIMIT: The DeepFake-TIMIT dataset [15] includes 640 DeepFake videos generated with faceswap-GAN [1] and based on the Vid-TIMIT dataset [28]. The videos are divided into two equal-sized subsets: DF-TIMIT-LQ and DF-TIMIT-HQ, with synthesized faces of size 64×64 and 128×128 pixels, respectively.

FF-DF: The FaceForensics++ dataset [25] includes a subset of DeepFakes videos, which has 1,000 real YouTube videos and the same number of synthetic videos generated using faceswap [2].

DFD: The Google/Jigsaw DeepFake detection dataset [10] has 3,068 DeepFake videos generated based on 363 original videos of 28 consented individuals of various genders, ages and ethnic groups. The details of the synthesis algorithm are not disclosed, but it is likely to be an improved implementation of the basic DeepFake maker algorithm.

DFDC: The Facebook DeepFake detection challenge dataset [9] is part of the DeepFake detection challenge, which has 4,113 DeepFake videos created based on 1,131 original videos of 66 consented individuals of various genders, ages and ethnic groups. This dataset is created using two different synthesis algorithms, but the details of the synthesis algorithm are not disclosed.

Based on release time and synthesis algorithms, we categorize UADFV, DF-TIMIT, and FF-DF as the *first generation* of DeepFake datasets, while DFD, DFDC, and the proposed Celeb-DF datasets are the *second generation*. In general, the second generation datasets improve in both quantity and quality over the first generation.

3. The Celeb-DF Dataset

The Celeb-DF dataset is comprised of 590 real videos and 5,639 DeepFake videos (corresponding to over two million video frames). The average length of all videos is approximate 13 seconds with the standard frame rate of 30 frame-per-second. The real videos are chosen from publicly available YouTube videos, corresponding to interviews of 59 celebrities (a full list is given in the supplementary material) with a diverse distribution in their genders, ages, and ethnic groups. 56.8% subjects in the real videos are male, and 43.2% are female. 8.5% are of age 60 and above, 30.5% are between 50 - 60, 26.6% are 40s, 28.0% are 30s, and 6.4% are younger than 30. 5.1% are Asians, 6.8% are African Americans and 88.1% are Caucasians. In addition, the real videos exhibit large range of changes in aspects such as the subjects' face sizes (in pixels), orientations, lighting conditions, and backgrounds. The DeepFake videos are generated by swapping faces for each pair of the 59 subjects. The final videos are in MPEG4.0 format. A comparison of the Celeb-DF dataset with other existing DeepFake datasets is summarized in Table 1.

4. Evaluating DeepFake Detection Methods

Using Celeb-DF and other existing DeepFake datasets, we perform the most comprehensive performance evaluation of DeepFake detection to date, with the largest number of DeepFake detection methods and datasets considered. There are two purposes of this evaluation. First, using the average detection performance as an indicator of the challenge levels of various DeepFake datasets, we further compare Celeb-DF with existing DeepFake datasets. Furthermore, we survey the performance of the current DeepFake detection methods on a large diversity of DeepFake videos, in particular, the high-quality ones in Celeb-DF.

4.1. Compared DeepFake Detection Methods

We consider nine DeepFake detection methods in our experiments. Because of the need to run each method on the Celeb-DF dataset, we choose only those that have code and the corresponding DNN-model publicly available or obtained from the authors directly.

- **Two-stream** [33] uses a two-stream CNN to achieve state-of-the-art performance in general-purpose image forgery detection. The underlying CNN is the GoogLeNet InceptionV3 model [30] trained on the SwapMe dataset [33]. We use it as a baseline to compare other dedicated DeepFake detection methods.
- **MesoNet** [4] is a CNN-based DeepFake detection method targeting on the mesoscopic properties of images. The model is trained on unpublished DeepFake datasets collected by the authors. We evaluate two variants of MesoNet, namely, *Meso4* and *MesoIncep*.

Methods	Model Type	Training Dataset	Repositories	Release Date
Two-stream [33]	GoogLeNet InceptionV3 [30]	SwapMe [33]	Unpublished code provided by the authors	2018.03
MesoNet [4]	Designed CNN	Unpublished	https://github.com/DariusAf/MesoNet	2018.09
HeadPose [32]	SVM	UADFV [32]	https://bitbucket.org/ericyang3721/headpose_forensic/	2018.11
FWA [17]	ResNet-50 [14]	Unpublished	https://github.com/danmohaha/CVPRW2019_Face_Artifacts	2018.11
VA-MLP [20]	Designed CNN	Unpublished	https://github.com/FalkoMatern/Exploiting-Visual-Artifacts	2019.01
VA-LogReg [20]	Logistic Regression Model			
Xception [25]	XceptionNet [8]	FaceForensics++ [25]	https://github.com/ondyari/FaceForensics	2019.01
Multi-task [21]	Designed CNN	FaceForensics [24]	https://github.com/nii-yamagishilab/ClassNSeg	2019.06
Capsule [23]	Designed CapsuleNet [27]	FaceForensics++	https://github.com/nii-yamagishilab/Capsule-Forensics-v2	2019.10
DSP-FWA	SPPNet [13]	Unpublished	https://github.com/danmohaha/DSP-FWA	2019.11

Table 2. Summary of compared DeepFake detection methods. See texts for more details.

tion4. Meso4 uses conventional convolutional layers, while MesoInception4 is based on the more sophisticated Inception modules [31].

- **HeadPose** [32] detects DeepFake videos using the inconsistencies in the head poses of the synthesized videos, based on a SVM model on estimated 3D head orientations from each video. The SVM model in this method is trained on the UADFV dataset.
- **FWA** [17] detects DeepFake videos using a ResNet-50 [14] to expose the face warping artifacts introduced by the resizing and interpolation operations in the basic DeepFake maker algorithm. This model is trained on self-collected face images.
- **VA** [20] is a recent DeepFake detection method based on capturing visual artifacts in the eyes, teeth and facial contours of the synthesized faces. There are two variants of this method: VA-MLP is based on a multilayer feedforward neural network classifier, and VA-LogReg uses a simpler logistic regression model. These models are trained on unpublished dataset, of which real images are cropped from CelebA dataset [19] and the DeepFake videos are from YouTube.
- **Xception** [25] corresponds to the baseline DeepFake detection method trained on the FaceForensics++ dataset, which is based on the XceptionNet model [8]. There are three variants of Xception, namely, *Xception-raw*, *Xception-c23* and *Xception-c40*. These three variants differ in the compression format of their training data: *Xception-raw* are trained on raw videos, while *Xception-c23* and *Xception-c40* are trained on H.264 videos with medium (23) and high degrees (40) of compression, respectively.
- **Multi-task** [21] is another recent DeepFake detection method that uses a CNN model to simultaneously detect manipulated images and segment manipulated areas as a multi-task learning problem. This model is trained on the FaceForensics dataset [24]. For relevance, we only consider the detection performance.
- **Capsule** [23] uses capsule structures [27] based on a VGG19 [29] network as the backbone architecture for DeepFake classification. This model is trained on FaceForensics++ dataset.
- **DSP-FWA** is a recently further improved method based on FWA, which includes a spatial pyramid pool-

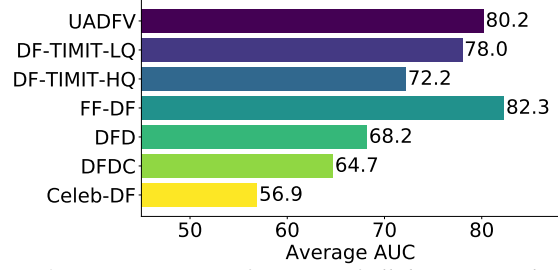


Figure 2. Average AUC performance of all detection methods on each dataset.

ing (SPP) module [13] to better handle the variations in the resolutions of the original target faces. This method is trained on self-collected face images.

A concise summary of the underlying model, source code, and training datasets of the DeepFake detection methods considered in our experiments is given in Table 2.

4.2. Experimental Settings

We evaluate the overall detection performance using the area under ROC curve (AUC) score at the frame level for all key frames. There are several reasons for this choice. First, all compared methods analyze individual frames (usually key frames of a video) and output a classification score for each frame. Using frame-level AUC thus avoids differences caused by different approaches to aggregating frame-level scores for each video. Second, using frame level AUC score obviates the necessity of calibrating the classification outputs of these methods across different datasets. To increase robustness to numerical imprecision, the classification scores are rounded to five digits after the decimal point, i.e., with a precision of 10^{-5} .

We compare performance of each detection method using the inference code and the published pre-trained models. This is because most of these methods do not have published code for training the machine learning models. As such, we could not practically re-train these models on all datasets we considered. We use the default parameters provided with each compared detection method.

4.3. Results and Analysis

In Table 3 we list individual frame-level AUC scores of all compared DeepFake detection methods over all datasets

Methods↓ Datasets→	UADFV [32]	DF-TIMIT [15]		FF-DF [25]	DFD [10]	DFDC [9]	Celeb-DF
		LQ	HQ				
Two-stream [33]	85.1	83.5	73.5	70.1	52.8	61.4	53.8
Meso4 [4]	84.3	87.8	68.4	84.7	76.0	75.3	54.8
MesoInception4	82.1	80.4	62.7	83.0	75.9	73.2	53.6
HeadPose [32]	89.0	55.1	53.2	47.3	56.1	55.9	54.6
FWA [17]	97.4	99.9	93.2	80.1	74.3	72.7	56.9
VA-MLP [20]	70.2	61.4	62.1	66.4	69.1	61.9	55.0
VA-LogReg	54.0	77.0	77.3	78.0	77.2	66.2	55.1
Xception-raw [25]	80.4	56.7	54.0	99.7	53.9	49.9	48.2
Xception-c23	91.2	95.9	94.4	99.7	85.9	72.2	65.3
Xception-c40	83.6	75.8	70.5	95.5	65.8	69.7	65.5
Multi-task [21]	65.8	62.2	55.3	76.3	54.1	53.6	54.3
Capsule [23]	61.3	78.4	74.4	96.6	64.0	53.3	57.5
DSP-FWA	97.7	99.9	99.7	93.0	81.1	75.5	64.6

Table 3. Frame-level AUC scores (%) of various methods on compared datasets. Bold faces correspond to the top performance.

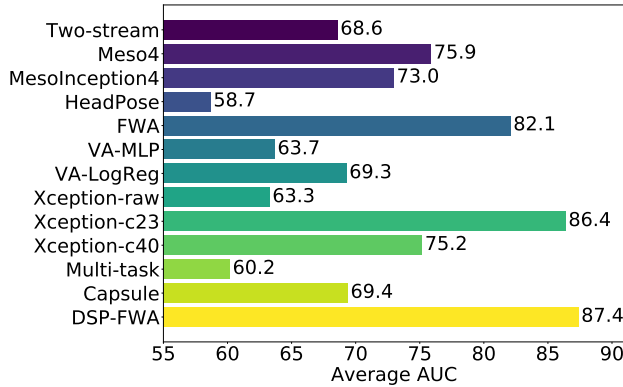


Figure 3. Average AUC performance of each detection method on all evaluated datasets.

including Celeb-DF.

Comparing different datasets, in Fig.2, we show the average frame-level AUC scores of all compared detection methods on each dataset. Celeb-DF is in general the most challenging to the current detection methods, and their overall performance on Celeb-DF is lowest across all datasets. These results are consistent with the differences in visual quality. Note many current detection methods predicate on visual artifacts such as low resolution and color mismatch, which are improved in synthesis algorithm for the Celeb-DF dataset. Furthermore, the difficulty level for detection is clearly higher for the second generation datasets (DFD, DFDC, and Celeb-DF, with average AUC scores lower than 70%), while some detection methods achieve near perfect detection on the first generation datasets (UADFV, DF-TIMIT, and FF-DF, with average AUC scores around 80%).

In term of individual detection methods, Fig.3 shows the comparison of average AUC score of each detection method on all DeepFake datasets. These results show that detection has also made progress with the most recent DSP-FWA

method achieves the overall top performance (87.4%).

5. Conclusion

In this work, we present a new challenging large-scale dataset for the development and evaluation of DeepFake detection methods. By using an improved synthesis algorithm, the Celeb-DF dataset brings closer the gap in visual quality of DeepFake datasets and the actual DeepFake videos circulated online. Based on the Celeb-DF dataset, we perform a comprehensive evaluation of the performance of current DeepFake detection methods, showing that there is still much room for improvement.

References

- [1] faceswap-GAN github. <https://github.com/shaoanlu/faceswap-GAN>, Accessed Nov 4, 2019.
- [2] faceswap github. <https://github.com/deepfakes/faceswap>, Accessed Nov 4, 2019.
- [3] FakeApp. <https://www.malavida.com/en/soft/fakeapp/>, Accessed Nov 4, 2019.
- [4] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.
- [5] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.
- [6] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder-decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing (TIP)*, 2019.
- [7] Robert Chesney and Danielle Keats Citron. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *107 California Law Review (2019, Forthcoming); U of Texas Law, Public Law Research Paper No. 692; U of Maryland Legal Studies Research Paper No. 2018-21*, 2018.
- [8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

- [9] Brian Dolhansky, Russ Howes, Ben Pfau, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.
- [10] Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. Deepfakes detection dataset by google & jigsaw.
- [11] Hany Farid. *Digital Image Forensics*. MIT Press, 2012.
- [12] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *AVSS*, 2018.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2015.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [15] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.
- [16] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.
- [17] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.
- [18] Yaqi Liu, Qingxiao Guan, Xianfeng Zhao, and Yun Cao. Image forgery localization based on multi-scale convolutional neural networks. In *ACM Workshop on Information Hiding and Multimedia Security (IHMMSec)*, 2018.
- [19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.
- [20] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019.
- [21] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2019.
- [22] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [23] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*, 2019.
- [24] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.
- [25] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.
- [26] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent-convolution approach to deepfake detection-state-of-art results on faceforensics++. *arXiv preprint arXiv:1905.00582*, 2019.
- [27] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *NeurIPS*, 2017.
- [28] Conrad Sanderson and Brian C Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *International Conference on Biometrics*, 2009.
- [29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.
- [31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.
- [32] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [33] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.
- [34] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *CVPR*, 2018.