On the Generality of Facial Forgery Detection

Joshua Brockschmidt
Computer Science and Engineering
University of Washington
Seattle, WA, USA
jcbrock@uw.edu

Jiacheng Shang and Jie Wu
Computer and Information Sciences
Temple University
Philedelphia, PA, USA
jiacshang@gmail.com and jiewu13@gmail.com

***Abstract*—A variety of architectures have been designed or repurposed for the task of facial forgery detection. While many of these designs have seen great success, they largely fail to address challenges these models may face in practice. A major challenge is posed by generality, wherein models must be prepared to perform in a variety of domains. In this paper, we investigate the ability of state-of-the-art facial forgery detection architectures to generalize. We first propose two criteria for generality: reliably detecting multiple spoofing techniques and reliably detecting unseen spoofing techniques. We then devise experiments which measure how a given architecture performs against these criteria. Our analysis focuses on two state-of-the-art facial forgery detection architectures, MesoNet and XceptionNet, both being convolutional neural networks (CNNs). Our experiments use samples from six state-of-the-art facial forgery techniques: Deepfakes, Face2Face, FaceSwap, GANnotation, ICface, and X2Face. We find MesoNet and XceptionNet show potential to generalize to multiple spoofing techniques but with a slight trade-off in accuracy, and largely fail against unseen techniques. We loosely extrapolate these results to similar CNN architectures and emphasize the need for better architectures to meet the challenges of generality.**

***Index Terms*—CNN, facial forgery detection, image forgery detection, video streaming**

I. Introduction

Online video streaming has become an integral channel of communication and information for much of the world's population. Video stream sites like YouTube and Vimeo find themselves at the center of a this vast exchange. Just recently in May of 2019, YouTube reached a monthly user base of two billion people, more than a quarter of the world's population [1]. Not only do these sites serves as entertainment, but as an integral means of staying informed about the world. In 2016, a survey found that 62% of U.S. adults getting their news from social media [2], a figure that is likely to be reflected in other technologically advanced countries.

Recently, the phenomena known as "deep fakes" has presented a significant challenge to the trustworthiness of digital videos. The term "deep fakes" refers generally to artificial intelligence-based techniques for convincingly manipulating faces, such that an individual can be made to appear as though they are saying or doing something they never did. Highlighting the dangers these technologies present, the US House Intelligence Committee recently held a hearing to discuss the national security concerns of deep fakes [3]. The discussion centered around the concept of "disinformation wars", a new form of cyberwarfare where spoofed videos could be used to interfere with elections and upset markets, and in turn erode public trust in journalism and the media. Outside the political landscape, the rise of "virtual influencers" hints at the potential for deep fakes to be used to mislead consumers [4]. These computer-generated models are being used to market products with the attention they gain from appearing to be real people. As tools like FaceSwap [5], DeepFaceLab [6], and the Deepfakes app [7] demonstrate, the tools for manipulating and spoofing human faces are increasing in availability, ease of use, and believability. The need to develop tools to aid social media sites and individuals in discerning real faces from fake faces is of increasing relevance.

This research was funded by the National Science Foundation as part of their Research Experiences for Undergraduates program.

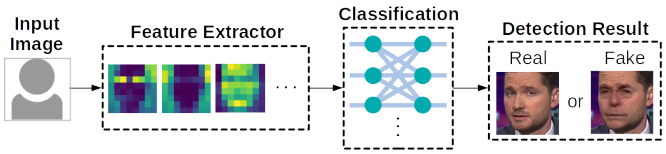
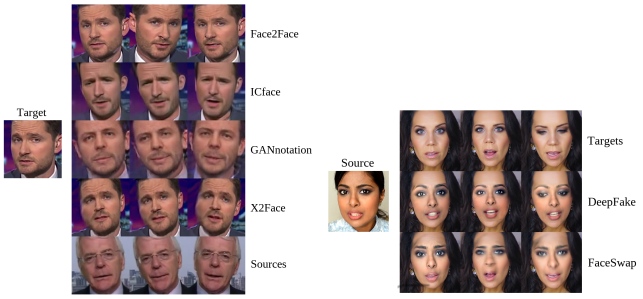


Fig. 1: Feature extractor and classifier.

In this paper, we are looking at forgery detection techniques built on CNNs. CNNs are specialized neural networks built to mimic the visual cortex, and are used extensively for image analysis. In particular, we are looking at CNNs used as simple binary classifiers, outputting a simple "real" or "fake" identification for a given image or video. These binary classifier can be thought of as containing two parts, a *feature extractor* and a *classifier*, pictured in Fig.1. The feature extractor is trained to extract features from a input image, which then used by the classifier to determine if the image is real or fake.

A number of architectures have been proposed for and applied to facial forgery detection [8] [9] [23] [22] [24] [21] [20]. Largely absent or sparsely mentioned in their analyses is their ability to detect multiple spoofing techniques simultaneously and the ability to detect unseen techniques. We refer to this as the problem of *generality*. Our analysis focuses on two CNNs, XceptionNet [8] and MesoNet [9]. These networks have been shown to outperform other similar CNNs in facial forgery detection [10], so we treat their behavior as exemplary of similar architectures. Our training and testing sets will consist of fake faces generated by the forgery methods of Deepfakes [7], Face2Face [11], FaceSwap [5], GANnotation [12], ICface [13], and X2Face [14]. These spoofing techniques are chosen



(a) Facial reenactment. (b) Face swapping.

Fig. 2: Facial forgery examples.

as they represent the state-of-the-art for various methods of face spoofing.

We perform three experiments to ascertain the potential of MesoNet and XceptionNet to generalize. For these experiments, we extend the FaceForensics++ dataset [10]. Alongside its video samples for Deepfakes, Face2Face, and FaceSwap, we generate analogous sequences with GANnotation, ICface, and X2Face. Our first experiment looks for similarities in the features extracted for models trained on different classes of fake faces. We find that while most classes shared identifying features, they have unique features which are necessary for higher detection accuracies, presenting a considerable challenge to generality. Our second experiment looks at how MesoNet and XceptionNet perform on unseen data. Both are revealed to perform quite poorly on unseen forgery techniques, with only a marginal improvement when grouping together multiple classes.

Our contributions are as follows:

- Present a new dataset of fake face videos generated with the generative adversarial networks (GANs) of GANnotation, ICface, and X2face.
- Demonstrate that simple CNN binary classifiers for facial forgery detection suffer a degradation in performance when trained on multiple forgery techniques.
- Show that simple CNN binary classifiers for facial forgery detection methods perform poorly against unseen forgery techniques.

The remainder of the paper is organized as follows: Section II provides preliminary knowledge of several facial forgery and detection methods. Section III describes three experiments, followed by an explanation of our dataset in Section IV. Each step of our experiments are described in Section V and the results discussed in Section VI. Finally, Section VII and VIII discuss related and future works, with our conclusion in Section IX

II. Preliminary

A. Facial Forgery Methods

We differentiate between two categories of facial forgery. The terms "source" and "target" here refer to faces in videos or images, where a source contains characteristics that will be transfer to a target:

- *Face swapping* - Transferring a face from a source onto a target, preserving the facial expression and pose of the target (see Fig. 2b).
- *Facial reenactment* - Transferring the facial expression or pose of a source onto a target, preserving the identity of the target (see Fig. 2a).

The term "deep fakes" has been used colloquially to refer to a wide range of face swapping techniques that utilize deep learning. However within the scope of this paper, *Deepfakes* [7] refers to a particular face swapping application. This AI-based technique trains a model to reconstruct images of a source and target face, then applies the portion of the model that reconstructs the source's face to the target's face to perform face swapping. *FaceSwap* [5] on the other hand is a more traditional graphics-based approach to face swapping. It uses facial landmarks to fit a 3D face model of a source face, which is then aligned with a target's face and blended with the original image. *Face2Face* [11] is a graphics-based approach to facial reenactment. It constructs a 3D model of a source face, which is then aligned with a target's face and the expressions transferred. To create videos, these processes are simply repeated frame-by-frame.

A more recent approach to facial reenactment is the use of GANs [15]. The general approach is to train a generator network to modify a source face to match the facial attributes of a target, and to in turn train a discriminator network to differentiate a real face from a fake face. By having these models compete, the generative model approaches photo-realistic results. Two such methods, GANnotation [12] and ICface [13], work by extracting facial attributes—like facial landmarks, head pose, or Action Units [16]—from a source image and transferring them to a target image. X2Face [14], another GAN-based method, uses an arbitrary driving vector to control a target face. This driving vector can be anything from audio to the same facial attributes used for GANnotation and ICface.

B. Facial Forgery Detection

This paper's analysis is limited to CNN architectures which perform the singular task of classifying individual images of faces as real or fake. There are other architectures that perform multiple tasks [23], are designed to adapt to new problem domains [22], or which look at an entire video rather than single frames [24]. But our analysis is limited to the aforementioned category of architectures because they are more prevalent than their more complex and specialized counterparts allowing us to draw more salient comparisons. One such architecture, **MesoNet**, proposed by Afchar et al. [9] refers to two CNNs designed specifically for facial forgery detection. They aim to overcome the data degradation introduced by video compression by focusing on the mesoscopic properties of images. We are focusing on MesoNet's second network, MesoInception-4, which uses a variant of Inception modules [17] to increase the range of features a model can extract. **XceptionNet**, proposed by Chollet [8] is an general image classification network derived from the Inception architecture

[17], where Inception modules have been replaced with depthwise separable convolutions to achieve similar effects. XceptionNet has been applied to facial forgery detection by Rössler et al. alongside their FaceForensics [18] and FaceForensics++ [10] datasets, where it has been shown to outperform MesoNet in the latter.

III. Overview

A. Adversary Model

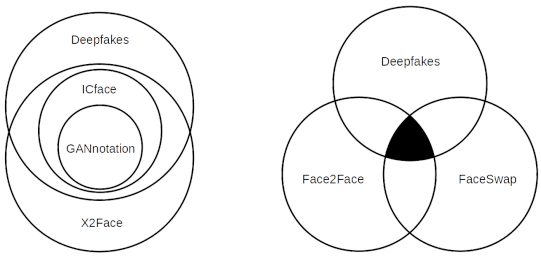
In our adversary model, a malicious user aims to use facial forgery to create videos where a victim appears to be saying or doing something they did not do. This video is then presented to an audience of unsuspecting viewers with the aim of spreading false information about the victim. We assume following: 1) the attacker has sufficient facial data of the victim to create a convincing fake video; 2) the attacker has sufficient time and resources to generate the fake video; 3) viewers of the fake video cannot visually identify it as fake. This presents a scenario where viewers will believe the contents of a spoofed video to be real if not assisted. The objective of facial forgery detection is to differentiate an attacker's video from real videos.

B. Experiment Configuration

We devise three sets of experiments to determine the generality of MesoNet and XceptionNet and in turn extrapolate the generality of similar architectures. Through our experiments we explore properties we will refer to as *feature overlap* and *transferability*. When a feature extractor for a class of fake images is found to extract features that can be used to identify another class of fakes, we say the original class has *feature overlap* with the second. *Transferability* refers to how well a model performs on fake classes it has not been trained on. Both these properties are considered to vary on a spectrum from low to high.

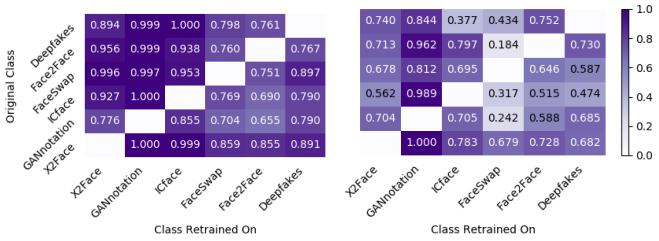
1) Feature Overlap: For models to generalize, their feature extractors must be capable of extracting identifying features for multiple forgery techniques. To this end, we investigate how much feature overlap their is between classes. More specifically, we use transfer learning on pretrained MesoNet and XceptionNet models, wherein we retrain models for one fake class on another fake class without modifying their feature extractors. The performance of these new models against their new fake classes will tell us how well the features extracted for the original fake classes overlap with the unseen classes'. What we are specifically interested in seeing are the properties of feature overlap and the overall degrees of feature overlap for MesoNet and XceptionNet and how they differ.

2) Transferability: We also want to see how well MesoNet and XceptionNet perform on unseen methods as we cannot depend on having knowledge of all attackers' methods. Most interesting to us is the overall transferability of MesoNet and Xception and how they differ.



(a) (b)

Fig. 3: Some feature overlaps for MesoNet.



(a) MesoInception-4 coefficients. (b) Xception coefficients.

Fig. 4: Feature overlap coefficients.

IV. Dataset

For our dataset, we extended FaceForensics++ [10] with the GAN-based reenactment techniques GANnotation, ICface, and X2Face. The original dataset consists of $1,000$ original video sequences and $3,000$ videos which have been manipulated by Face2Face, Deepfakes, and FaceSwap, with predefined training, testing, and validation splits. All original and manipulated sequences are provided on three levels of compression using the H.264 codec with quantizations of 0, 23, and 40, which we will refer to as lossless, visibly lossless, and lossy, respectively. Using the same sources and targets, we generated $3,000$ additional videos with GANnotation, ICface, and X2Face and their compressed counterparts. For all $7,000$ videos and for all three levels of compression, we extracted a cropped image of a face every 30 frames. $293,975$ image samples were extracted in total.

V. Experiments

All our models are trained with the ADAM optimizer [19] with a learning rate of 0.001, betas of 0.9 and 0.999, an epsilon of 10^7, and use samples from all three compression levels.

A. Feature Overlap

We started with MesoInception-4 and Xception models trained on single facial forgery techniques for all compression levels. For each model we froze their convolutional layers, reset the neurons in their classification layers, and retrained them on another forgery class for all combinations of classes. To measure the performance of our new model, we devise the *feature overlap coefficient*,

$$f(x, y) = \frac{z_y}{z_x} \tag{1}$$

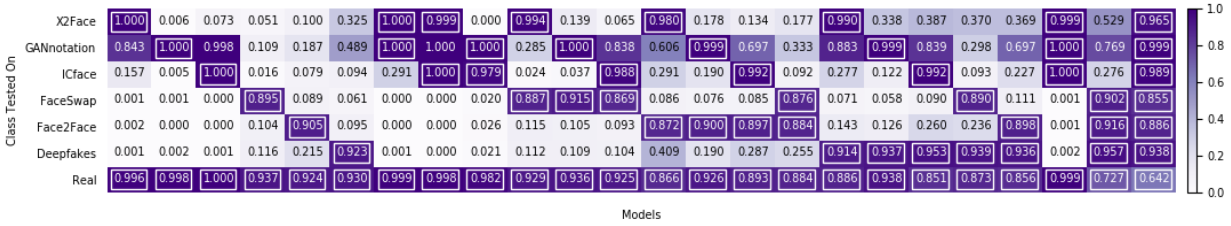


Fig. 5: Accuracies for Xception against all classes. Each column is a model with the trained classes outlined.

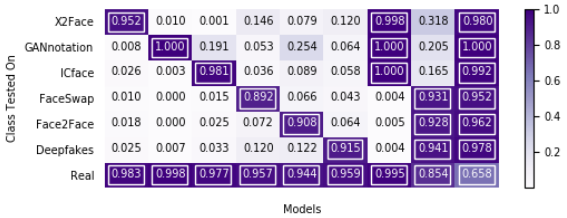


Fig. 6: Accuracies for Xception against all classes. Each column is a model with the trained classes outlined.

where x is the class the feature extractor belongs to, y is the class transferred to, z_x is the true detection rate for class x with the original model, and z_y is the true detection rate for class y with the retrained model. We suppose a higher value of f indicates a higher degree of feature overlap of x with y, where a value of 1 or greater is ideal. Our resulting coefficients are plotted in Fig. 4a for MesoInception-4 and Fig. 4b for Xception, where the rows correspond to the class transferred from and columns to the class transferred to.

B. Transferability

For this experiment, we trained MesoInception-4 and Xception models on different combinations of fake classes then tested them against all classes. When training on two or more fake classes, there is a considerable imbalance between real and fake samples. To combat this, we had the loss function weigh every real and fake sample by the reciprocal of the total number of real and fake in use, respectively. The resulting accuracies are shown in Fig. 5 and Fig. 6. Higher accuracies on unseen classes indicate higher transferability.

VI. Discussion

A. Feature Overlap

Comparing MesoInception-4 to Xception in Fig. 4, we see that MesoInception-4 has more feature overlap. This tell us that degrees of feature overlap are not consistent across architectures. Likewise, we see different *relative* degrees of feature overlap between classes. For MesoInception-4, for example, there is more feature overlap for Deepfakes to ICface than Deepfakes to X2Face, whereas the opposite is true for Xception. To highlight some important properties of feature overlap, we visualize several feature overlap relationships for MesoInception-4 in Fig. 3 as Venn diagrams. An important takeaway from Fig. 3a is that while classes like GANnotation and ICface are fully encompassed by Deepfakes' features, the opposite is not true. Feature overlap is rarely a one-to-one relationship. An important limitation of our analysis is pictured in Fig. 3b. While we can ascertain there are shared features between Deepfakes and Face2Face, Deepfakes and FaceSwap, and Face2Face and FaceSwap, we cannot determine what degree of features all three share. This unknown three-way relationship is shaded black. We cannot expect feature overlap relationships to be entirely consistent across architectures, and some classes have more feature overlap than others. The inconsistent degrees of feature overlap among various fake classes for MesoInception-4 and Xception suggest they will have difficulties generalizing to multiple facial forgery techniques.

B. Transferability

Looking first at our MesoInception-4 models in Fig. 5, we see there is very low transferability. Very few unseen classes achieve accuracies above 20%, and GANnotation is the only class to achieve decent unseen accuracies. Increasing the number of classes trained on does appear to increase transferability. But this comes at the cost of lower real detection rates. Looking at our results for Xception in Fig. 6, we see even less transferability. Like MesoInception-4, the transferability for Xception increases as we train on more classes, but at the expense of real detection accuracy. Overall, both MesoInception-4 and Xception perform very poorly against unseen facial forgery techniques.

VII. Related Work

A number of architectures that mimic the essential CNN structure of MesoNet and Xception shown in Fig. 1 have been proposed for image forgery detection. Rahmouni et al. [20] use "patch classification" to distinguish photo-realistic computer graphics from natural images. They split an image into tiles and tally to probability of each being fake to arrive at a final verdict. Bayar and Stamm [21] proposed a CNN for general image forgery detection that ignores the content of images. Their aim is to focus on pixel-to-pixel relationship that distinguish artifacts from different forgery techniques. These networks have achieved accuracies of $44\% - 70\%$ and $68\% - 88\%$, respectively, against Deepfakes, Face2Face, and FaceSwap [10]. In contrast to these networks are several recent CNN architectures that expand upon the basic feature extraction and classification structure. ForensicTransfer, proposed by Cozzolino et al. [22], utilizes transfer learning to increase transferability. Given only a only a few samples of an unseen forgery technique, the network can be retrained to

effectively classify it. Nguyen et al. [23] proposed a multi-task learning that both detects facial forgery and segments the manipulated region of an image simultaneously. These tasks share information with the aim of improving overall performance and transferability. Sabir et al. [24] proposed a recurrent network that exploits temporal information in videos rather than looking at single frames in isolation.

VIII. Future Works

Similar experiments could be performed on more advanced convolutional architectures like ForensicTransfer [22] or Nguyen et al. [23]. While these detection methods are designed to accomodate transferability, a more in-depth analysis could reveal why their methods succeed and where they need improvement. Our findings could also be used to inform the creation of more generalizable models or counter-detection efforts. Lastly, an exploration of more efficient architectures and how improved run times correlate with generality could use a similar analysis as ours. Designing facial forgery detection models that are both efficient and general will be crucial both for processing large streams of data on video streaming sites and for integrating with consumer devices.

IX. Conclusion

In this paper, we analyze the ability of facial forgery detection to generalize. We look at two state-of-the-art CNN architectures for detecting facial forgery, MesoNet and Xception, and their ability to generalize. We devised quantitative methods for ascertaining feature similarities (and dissimilarities) between models trained on different techniques and measuring transferability. With these methods we found that both architectures are capable of achieving consistent accuracies across varying compression levels without significant sacrifices in accuracy, but largely fail when tested against unseen data. While networks of this type show potential to generalize, they ultimately fail to accurately and reliably detect unseen methods. We must explore new architectures to achieve truly general facial forgery detection.

Acknowledgment

This work was supported by the NSF grant CNS-1757533 as part of the Research Experiences for Undergraduates (REU) program. Research was facilitated by Temple University during the Pervasive Computing for Smart Health, Safety, and Well-being REU.

References

[1] "Global logged-in YouTube viewers 2019," Statista. [Online]. Available: https://www.statista.com/statistics/859829/logged-in-youtube-viewers-worldwide/. [Accessed: 28-Jul-2019].

[2] J. Gottfried and E. Shearer, "News Use Across Social Media Platforms 2016," 26-May-2016. [Online]. Available: https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/. [Accessed: 28-Jul-2019].

[3] "House holds hearing on deepfakes and artificial intelligence amid national security concerns." [Online]. Available: https://www.cbsnews.com/news/house-holds-hearing-on-deepfakes-and-artificial-intelligence-amid-national-security-concerns-live-stream/. [Accessed: 26-Jul-2019].

[4] D. Fowler, "The fascinating world of Instagrams virtual celebrities." [Online]. Available: http://www.bbc.com/worklife/article/20180402-the-fascinating-world-of-instagrams-virtual-celebrities. [Accessed: 26-Jul-2019].

[5] "FaceSwap." [Online]. https://github.com/MarekKowalski/FaceSwap

[6] "DeepFaceLab." [Online]. https://github.com/iperov/DeepFaceLab

[7] "Deepfakes." [Online]. https://github.com/deepfakes/faceswap

[8] F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 12511258.

[9] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 17.

[10] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," arXiv:1901.08971 [cs], Jan. 2019.

[11] J. Thies, M. Zollhfer, M. Stamminger, C. Theobalt, and M. Niener, "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 23872395.

[12] E. Sanchez and M. Valstar, "Triple consistency loss for pairing distributions in GAN-based face synthesis," arXiv:1811.03492 [cs], Nov. 2018.

[13] S. Tripathy, J. Kannala, and E. Rahtu, "ICface: Interpretable and Controllable Face Reenactment Using GANs," arXiv:1904.01909 [cs], Apr. 2019.

[14] O. Wiles, A. Sophia Koepke, and A. Zisserman, "X2Face: A network for controlling face generation using images, audio, and pose codes," presented at the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 670686.

[15] I. Goodfellow et al., "Generative Adversarial Nets," in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 26722680.

[16] P. Ekman and W. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," Consulting Psychologists Press, 1978.

[17] C. Szegedy et al., "Going Deeper With Convolutions," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 19.

[18] A. Rssler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niener, "FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces," arXiv:1803.09179 [cs], Mar. 2018.

[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," ICML, 2014.

[20] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in 2017 IEEE Workshop on Information Forensics and Security (WIFS), 2017, pp. 16.

[21] B. Bayar and M. C. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, New York, NY, USA, 2016, pp. 510.

[22] D. Cozzolino, J. Thies, A. Rssler, C. Riess, M. Niener, and L. Verdoliva, "ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection," arXiv:1812.02510 [cs], Dec. 2018.

[23] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos," arXiv:1906.06876 [cs], Jun. 2019.

[24] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," arXiv:1905.00582 [cs], May 2019.