big data and cognitive computing

MDPI

Article

ADD: Attention-Based DeepFake Detection Approach

Aminollah Khormali and Jiann-Shiun Yuan *

Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816, USA; aminkhormali@knights.ucf.edu

* Correspondence: Jiann-Shiun.Yuan@ucf.edu; Tel.: +1-407-823-5719

Abstract: Recent advancements of Generative Adversarial Networks (GANs) pose emerging yet serious privacy risks threatening digital media's integrity and trustworthiness, specifically digital video, through synthesizing hyper-realistic images and videos, i.e., DeepFakes. The need for ascertaining the trustworthiness of digital media calls for automatic yet accurate DeepFake detection algorithms. This paper presents an attention-based DeepFake detection (ADD) method that exploits the fine-grained and spatial locality attributes of artificially synthesized videos for enhanced detection. ADD framework is composed of two main components including face close-up and face shut-off data augmentation methods and is applicable to any classifier based on convolutional neural network architecture. ADD first locates potentially manipulated areas of the input image to extract representative features. Second, the detection model is forced to pay more attention to these forgery regions in the decision-making process through a particular focus on interpreting the sample in the learning phase. ADD's performance is evaluated against two challenging datasets of DeepFake forensics, i.e., Celeb-DF (V2) and WildDeepFake. We demonstrated the generalization of ADD by evaluating four popular classifiers, namely VGGNet, ResNet, Xception, and MobileNet. The obtained results demonstrate that ADD can boost the detection performance of all four baseline classifiers significantly on both benchmark datasets. Particularly, ADD with ResNet backbone detects DeepFakes with more than 98.3% on Celeb-DF (V2), outperforming state-of-the-art DeepFake detection methods.

Citation: Khormali, A.; Yuan, J.-S. ADD: Attention-Based DeepFake Detection Approach. *Big Data Cogn. Comput.* **2021**, *5*, 49. https://doi.org/10.3390/bdcc5040049

Academic Editor: Min Chen

Received: 21 July 2021
Accepted: 26 August 2021
Published: 27 September 2021

Keywords: computer vision; cybersecurity; generative adversarial networks; DeepFake detection

1. Introduction

The recent advances in the field of deep learning, specifically generative adversarial networks [1,2] and convolutional auto-encoders [3], have significantly propelled the generation of sophisticated and compelling forged versions of misinformation of all kinds. Generally, fake information is carried out for malicious purposes, such as propaganda or misinformation campaigns. In the context of digital video, sophisticated image and video manipulation techniques have emerged as one of the most sinister forms of misinformation, posing emerging yet increasing privacy risks targeting large-scale communities all over the world [4–8]. Such manipulated videos are so sophisticated that they are extremely difficult to detect using state-of-the-art Artificial Intelligence (AI) visual media forensic tools, or even with human eyes [9]. Moreover, recent improvements in computer vision and deep learning techniques have made it extremely easy to create fake videos called DeepFakes, hyper-realistic and deceptive videos of real people by manipulating the face region while leaving only minimal visual artifacts [10,11]. Mainly, DeepFakes are the product of merging, combining, replacing, and superimposing images and videos using AI techniques to generate fake digital videos that appear authentic [12]. While initial DeepFake videos were benign and plain, created for fun or artistic values, adversaries abused this technology for malicious purposes leading to severe political, social, financial, and legal consequences [12–15]. The DeepFake videos' impact becomes more critical considering the scope, scale, and sophistication of the technology involved, as they can be fabricated using a simple computer [14]. Furthermore, DeepFake generation algorithms are evolving continually, which not only improve their visual quality but also makes them better at circumventing existing detection methods.

Thanks to the accessibility of large-volume training data, high-throughput computing power, and automated generation procedures, there has been a huge surge in developing new DeepFake creation algorithms. DeepFake generation methods can be categorized into before deep learning approaches [3,16–20] and deep learning-based approaches [21–25]. Despite small differences in the design of different DeepFake generators, they all follow the same flow. The common flow is to take in a video of a specific individual (target) and replace its face with another person (source). The backbone of the recent deceptive algorithms is generative adversarial networks, which map the source's facial expressions to the target through which it can achieve a high level of realism with a proper post-processing step [11].

As DeepFakes became super-realistic and more pervasive, ascertaining a digital video's trustworthiness and deciding on its authenticity becomes a more demanding yet challenging task. The fact that DeepFakes are created exploiting an AI algorithm rather than a camera capturing real events implies that they can still be detected using advanced deep learning networks [26]. Recently, multiple research works have focused on presenting a comprehensive understanding of the state-of-the-art methods and comparative analysis of DeepFakes [27–29]. The literature in this field shows that DeepFakes are inherently equipped with different artifacts ranging from visible artifacts as in earlier DeepFakes [11,30,31] to more hidden traces in more sophisticated DeepFakes [10,32,33], which can be exploited using high-level AI models to develop an automated digital video authentication system.

Objectives. This paper's primary goal is to present a digital video authentication system that offers high detection performance while covering a wide range of possible manipulation techniques. Such a digital media forensics tool is vital in the real-world scenario, considering the adversary's ever-evolving techniques in generating more deceptive DeepFakes. In general, training a new detection model is a computationally heavy and time-consuming process or even impractical due to a lack of sufficiently labeled data from the new manipulation technique. However, this goal can be achieved by forcing the model to learn hidden traces and intrinsic representations from manipulated regions.

Contributions. In this work, we look at the DeepFake detection task as a Fine-Grained Visual Classification (FGVC) problem. In both assignments, the main goal is to recognize the subordinate-level categories under a basic-level category. First, there is a substantial variance in the same class's images in terms of poses and viewpoints of the face, even for a person in the same video. Second, there is a minimal variance between the two different class images. The difference between the original and the fabricated image is tiny enough to deceive even human eyes, as can be observed in Figure 1. Furthermore, the forgery involves only the face region and leaves the background and other portions intact. By taking these characteristics into account, we developed a digital video authentication system, i.e., ADD, built based on an attention mechanism. ADD first locates potentially manipulated areas of the input image and extracts key representative features. Second, ADD forces the detection model to pay extra attention to these manipulated regions for decision making by imposing additional supervision on instance interpretation in the learning procedure through attention-based data augmentation. Finally, the performance of the ADD is evaluated against two challenging DeepFake forensic datasets. Comparing the obtained results with other existing models clearly demonstrate the excellence of the ADD in the given task. The major contributions of this work are summarized as follows:

- We considered the Deepfake detection task as an FGVC problem and proposed a digital video authentication system, ADD, built based on an attention mechanism.
- ADD first locates potentially manipulated areas of the input image and extracts discriminative features from those regions. Second, the detection model is made to pay more attention to these forgery regions for decision-making by imposing additional supervision on instance interpretation in the learning procedure through attention-based data augmentation.
- The performance of the ADD is evaluated against two challenging DeepFake forensic datasets. Experimental results demonstrate that ADD could achieve a detection rate of 98.37% on Celeb-DF (V2), outperforming state-of-the-art DeepFake detection methods.

Organization. The rest of the paper is organized as follows. In Section 2 we review the related works on DeepFake generation and detection techniques. In Section 3 we describe ADD outlining its three main components, including frame-wise face localization, localized discriminative features, and attention-based data augmentation. In Section 4 we review overall evaluation settings, including datasets, baseline network architectures, implementation specifics, and evaluation metrics. In Section 5 we discuss the experimental results of ADD, and finally, a conclusion is drawn in Section 6.

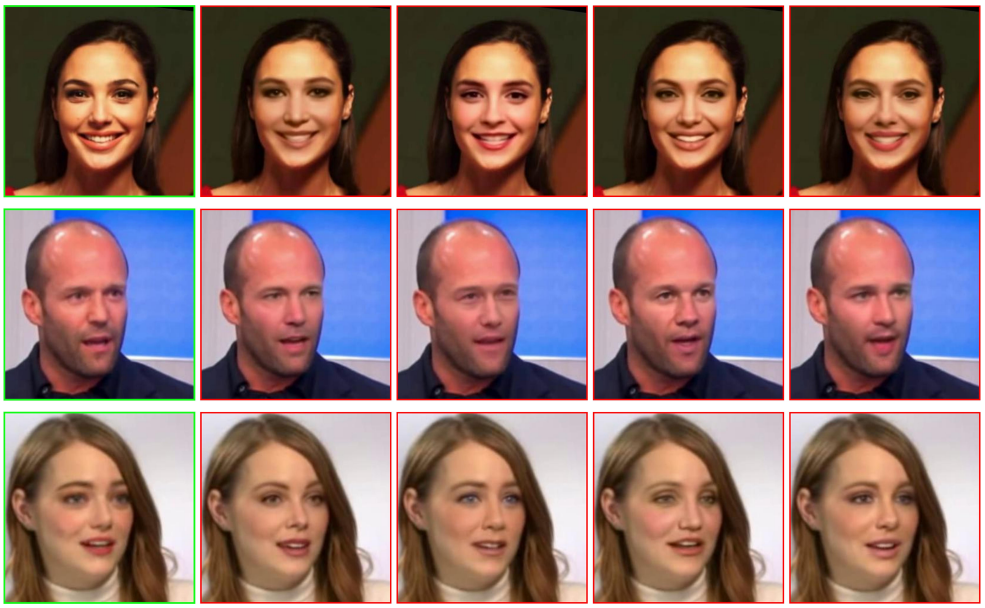

Figure 1. Example frames of DeepFake videos [25]. The left column (green border) is a selected frame from original videos, and other columns (red border) are corresponding AI-generated frames. Note, the intra-class variance is high, whereas the inter-class variance is small.

2. Related Work

In general, the field of DeepFake video analysis can be categorized into two broad domains, including DeepFake generation and DeepFake detection.

DeepFake Generation. Early DeepFake generation methods were simple and mostly relying on traditional vision and voice impersonation; however, most current methods involve sophisticated AI-based generation techniques, i.e., GANs. FakeApp was the first DeepFake creation software developed by a Reddit user using an autoencoder–decoder pairing structure [34,35]. Furthermore, Thies et al. [3] presented a real-time face capture and re-enactment of videos using a non-rigid model-based bundling. Masi et al. [19] presented a face-specific data augmentation technique using 3D shapes, and appearances of faces. Recent advancements in the field of deep learning have enabled adversaries to devise more sophisticated DeepFake creation techniques, leading to super-realistic videos, exploiting the unique generation capabilities of generative adversarial networks.

For example, Zhu et al. [36] and Kim et al. [37] have modified the GANs and presented cycle-consistent GANs to modify the domains of the output images based on the input image's domain. They have utilized this method for DeepFake generation where the source person's identities were changed to the target person while keeping the facial expression unchanged. Lu et al. [38] presented identity-guided conditional CycleGAN to create high-resolution face images from its low-resolution peers. Similarly, Kim et al. [23] presented deep video portraits that transfer both facial expression and 3D poses of the source image into the target image. Moreover, Faceswap-GAN [39] improves the visual quality of the synthesized images with adversarial and perceptual losses. The generated videos were more realistic thanks to the frame-to-frame face detection box's temporal smoothing and an attention mask. Thies et al. [16] presented a facial reenactment forgery method, NeuralTextures, based on a patch-based adversarial loss alongside a photometric reconstruction loss. Wang et al. [40,41] presented a flow-based face reenactment forgery

method known as the video-to-video synthesis approach based on multiple talking videos of the source and generating new DeepFakes using a single image of the target. In a similar approach, Siarohin et al. [42] incorporated a learnable optical flow network approximation to a first-order Taylor polynomial to generate a manipulated video of a person using a single image. Li et al. [43] presented Faceshifter, an adaptive attention-based denormalization generator for high-quality face replacement using a heuristic error acknowledging refinement network learning method.

DeepFake Detection. A large body of work in the DeepFake analysis domain is focused on devising automated yet effective detection techniques. Early detection techniques were focused on handcrafted features, i.e., blinking inconsistencies [11], biological signals [44], and unrealistic details [45]. Although manually crafted detection features helped to advance the DeepFake detection domain, their performance was poor and could be easily circumvented. Techniques based on deep learning networks are utilized lately to overcome this issue and build more reliable forgery detection tools. For instance, Afchar et al. [33] proposed the MesoNet that detects forgeries at an intermediate level of detail using a shallow convolutional network while avoiding microscopic features that can be eliminated during the video compression process. Cozzolino et al. [46] proposed the forensictransfer method, a forgery detection approach that is built based on autoencoder architecture and transfer learning. Nguyen et al. extended this method [47] by replacing the standard decoder with a decoder that generates a mask of the manipulated region using a multitask learning approach. Furthermore, Nguyen et al. [32] proposed Capsule-Forensics method to detect both replay attacks and digitally generated images and videos. Rana et al. [48] introduced a technique that combines a series of deep learning classification models and creates an improved composite classifier for DeepFake detection.

While the previously discussed approaches target intraframe dissimilarities, Güera and Delp [49] utilized time-distributed features and a long short-term memory network for DeepFake detection. Furthermore, Sabir et al. [50] evaluated the same approach using ResNet [51] and DenseNet [52] feature extractors, where the extracted faces were aligned in consecutive order using facial landmarks to maintain temporal consistency. Furthermore, Yu et al. [53] investigated the potential of GAN fingerprinting analysis for DeepFake detection. Dordevic et al. [54] presented a method based on scale-invariant feature transform for DeepFake detection. Kaur et al. [55] presented a sequential temporal analysis to detect face-swapped video clips using convolutional long short-term memory. Mittal et al. [56] presented an approach that simultaneously exploited audio and video modalities and perceived emotions from the two modalities for DeepFake detection.

Although researchers in the community have investigated the DeepFake detection problem from various perspectives, only minimal effort has been devoted to investigating DeepFakes from a fine-grained visual classification point of view, especially using attention-based techniques. The most similar works to ADD are [26,57] methods. In line with [26,57], our proposed method looks at the DeepFake detection problem as a fine-grained visual classification task while utilizing attention-based data augmentation techniques. However, our proposed method is different from [26] where the authors proposed a DeepFake detection method from FGVC angle that is built using an autoencoder structure different from our proposed method, which is based on a deep learning structure. Furthermore, ADD is different from [57] as ADD considers only the last two convolutional blocks in the model for data augmentation rather than the whole convolutional blocks, as it is proposed in [57]. Besides, ADD uses two different modules, i.e., Face close-up and Face Shut-off, to force the model to extract more discriminative information from different parts of face region; however, [57] generates attention masks focused on only eyes, nose, and mouth for adjusting the feature map of the face.

3. ADD: Methods

In this section, the proposed framework for the attention-based digital video authentication system, ADD, is introduced. The general pipeline of the presented attention-based DeepFake detection approach is illustrated in Figure 2. ADD which is composed of three main components, including face localization and preprocessing Section 3.1, local-

ized discriminative feature extraction Section 3.2, and attention-based data augmentation Section 3.3 followed by a classifier to distinguish original frames from DeepFakes.

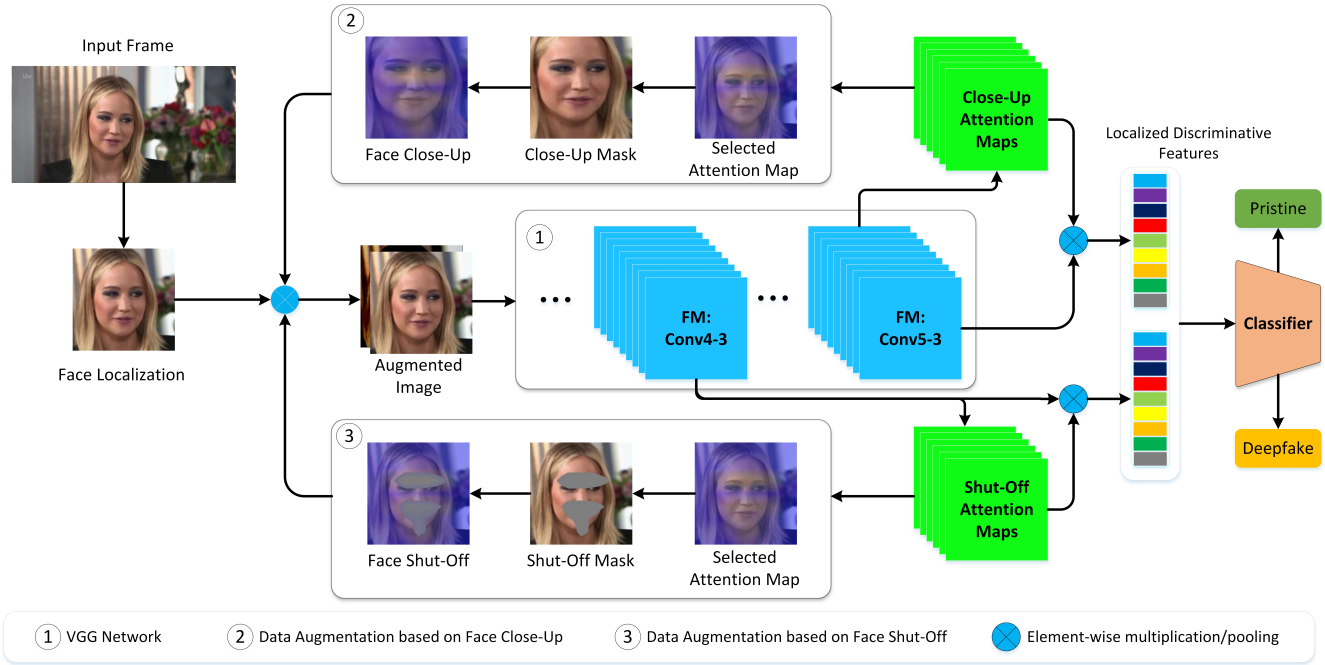

Figure 2. General structure of the proposed digital video authentication system. Representative features of each frame of a given video are calculated using face close-up and face shut-off attention mechanisms. While the face close-up attention mechanism enlarges and focuses on distinctive parts of the face, the face shut-off attention mechanism helps extracting other discriminative features from other parts of the image. Note that the classifier component of the ADD can be replaced with CNN-based classifiers.

3.1. Face Localization and Preprocessing

As it is pointed out in Section 1, DeepFake generation algorithms mainly manipulate face regions and leave the background part intact. Therefore, focusing on the face region of a video frame, instead of analyzing the whole frame as input to the learning model, not only improves the detection performance by reducing background noise but also reduces the computational time by reducing the size of the input sample [58]. To this end, the following steps, as shown in Figure 3, are taken for face localization and further analysis. First, for each input video, 20% of the frames are extracted in consecutive order, yielding to over 2 million frames on Celeb-DF (V2) dataset. Second, the state-of-the-art face detection method, i.e., RetinaFace [59], is utilized to locate facial landmarks on each extracted frame. The obtained facial landmarks are utilized to crop, align, and resize the faces to standard configuration [60]. These cropped frames, containing only face regions, are further used for attention-based image augmentation and feature extraction.

3.2. Localized Discriminative Features

In the DeepFake detection task, it is essential to determine the face region along with different facial landmarks for effective feature extraction. In this work, the distribution of face regions and associated facial landmarks is represented using attention maps. For a given frame I, the feature maps $F \in R^{H \times W \times C}$ are extracted using a CNN-based feature extractor, where H, W and C represent feature layer's height, width, and the number of channels, respectively. The obtained feature maps, F, are then utilized to calculate the distribution of M different parts of the face, i.e., Attention Maps $A \in R^{H \times W \times M}$, using a convolutional function $f(\cdot)$ as $A = f(F) = \bigcup_{k=1}^{M} A_k$. Here, each specific part of the face, i.e., lips, eyes, forehead, etc., are represented using $A_k \in R^{H \times W}$. Having generated M attention maps corresponding to M different parts of the face, representative feature maps of those parts F_k can be obtained by element-wise multiplication of feature maps F with

each attention map A_k. This process is shown in Figure 4. A feature extractor, e.g., global pooling function $g(\cdot)$, is utilized along with each of these local feature maps F_k to pool out more discriminative local features associated with k_{th} attention feature $f_k \in R^{1\times C}$. Finally, these local features f_k are stacked to build a comprehensive and distinctive feature set containing detailed information of the whole frame. Passing this valuable information to the model enforces the model to focus specifically on the forgery regions of a given input image and learn local interpretations to perform its decision-making.

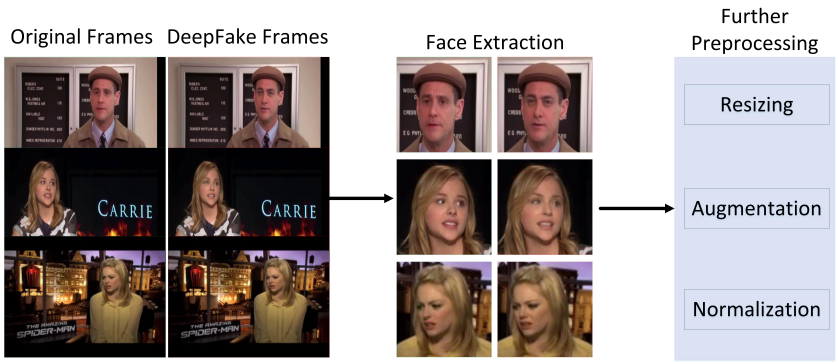

Figure 3. Frame-wise face localization pipeline. For every input video, 20% of the frames are extracted in consecutive order. For every obtained frame, facial landmarks are calculated using RetinaFace [59] to find and crop facial regions. Finally, all cropped faces are resized, augmented, and normalized for further analysis.

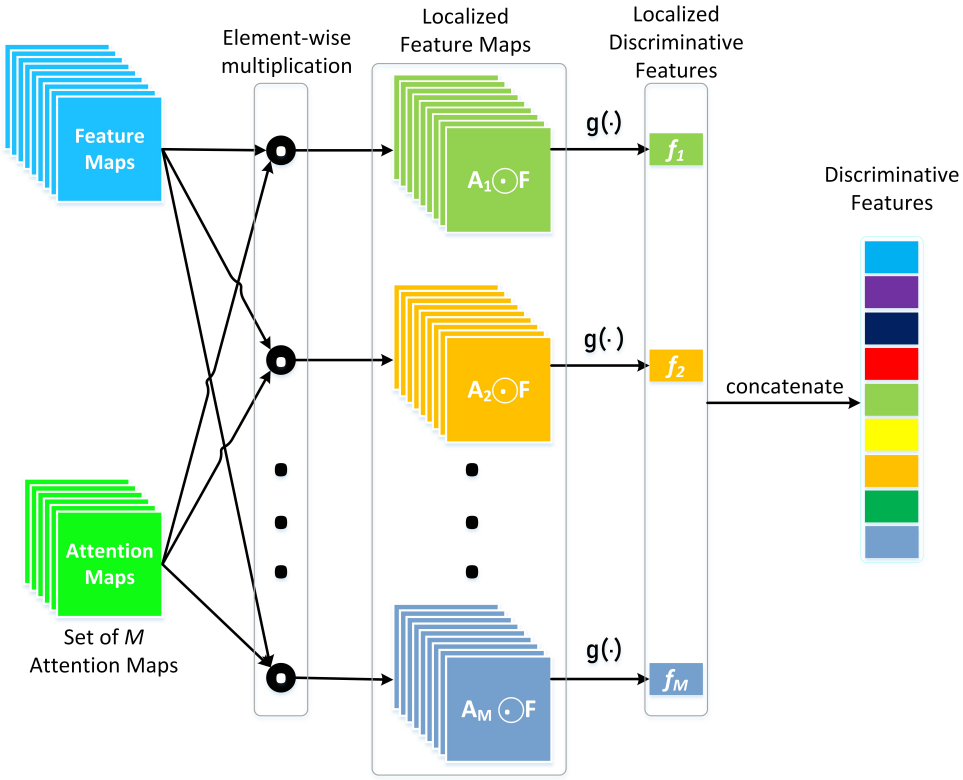

Figure 4. Localized discriminative feature extraction framework. Element-wise multiplication of feature maps with specific attention map results in localized feature maps to pool more discriminative features.

3.3. Attention-Based Data Augmentation

Once the attention maps are calculated, they can be employed for more efficient data augmentation. The problem with random data augmentation methods is their low effi-

ciency and introducing background noise. However, attention-based data augmentation is more efficient as it exposes the model to additional instance interpretation in the learning procedure. Two complimentary attention-based data augmentation approaches are employed in this work, including Face Close-Up and Face Shut-Off. While the former looks closer at a specific region of the face, the latter approach ignores that area and sees other face regions.

Face Close-Up. The face close-up augmentation approach's primary goal is to look closer at specific regions of the face, e.g., eyes, forehead, lips, etc., and provide more distinctive local features to the model to enhance its local interpretability. The following steps are taken into account to perform face close-up augmentation. One attention map is randomly selected from M available attention maps for each frame, and its elements are normalized to $[0, 1]$. All elements with a value greater than a particular predefined threshold are set to one, and the remaining are set to zero. Finally, only the region enclosed into a bounding box that covers all active areas is selected. The face close-up augmentation approach enlarges the scale of the face's selected region from raw input, thus improving the detection model's local explainability by focusing on the forgery region while being exposed to more fine-grained features. The augmented image is illustrated in Figure 2.

Face Shut-Off. While the face close-up approach provides a closer look into specific regions of the face, the resulting bounding boxes for different attention maps might be very similar. In such cases, the model would not learn new representative features. To avoid this issue and extract more discriminative features from other regions, the face shut-off data augmentation approach is utilized. Like the previous approach, for each frame, one attention map out of M available attention maps is randomly selected and normalized to $[0, 1]$. All the normalized attention map elements with a value greater than a particular predefined threshold are set to zero, whereas the remaining parts are set to one. This results in removing the active parts from the image, which in return forces the model to see other parts of the image and attain additional localized discriminative features.

4. Evaluation Settings

This section is devoted to introducing the overall evaluation settings, including the DeepFake detection datasets, baseline network architectures, implementation specifics, and evaluation metrics.

4.1. Datasets

To make a real-world impact and bear strong relevance of any digital video authentication system, it is crucial to evaluate the system against high-quality DeepFake datasets. The dataset should be super-realistic and stealthy while covering more diverse real-world scenes, and having minimal visual artifacts to maintains its high visual quality. Different research groups in the community have introduced different DeepFake detection datasets, such as UADFV dataset [61], the DeepFake-TIMIT dataset (DF-TIMIT) [62], the FaceForenscics++ dataset (FF-DF) [58], and the FaceBook DeepFake detection challenge (DFDC) dataset [63]. While this has considerably advanced the DeepFake detection in the early stages, most of them are far from perfect for today's real-world applications. They have major visual problems, such as limited scenes in original videos, low-quality synthesized faces, visible splicing boundaries, color mismatch, visible parts of the original face, and inconsistent synthesized face orientations [25,57].

Thus, in this study, the performance of the proposed method is empirically evaluated against two most recent and challenging DeepFake datasets, i.e., *Celeb-DF (V2)* [25] and *WildDeepfake* [57]. The former is a dataset with the highest visual quality score reported to date, and the latter is a challenging real-world DeepFake dataset with more diverse scenes and more persons with rich facial expressions in each scene.

Celeb-DF (V2). The Celeb-DF (V2) is a large-scale challenging video dataset of 590 original videos of celebrities and 5639 high-quality DeepFake videos generated using an improved synthesis process, corresponding to over 2 million frames. Real videos are collected from publicly available YouTube videos, and the fake ones are created by swapping faces for each pair of the subjects.

WildDeepfake. The WildDeepfake is a challenging real-world DeepFake detection dataset, where, unlike other datasets, both real and DeepFake videos are collected completely from the internet. This dataset presents more diverse scenes, more persons in each scene, and rich facial expressions. Corresponding dataset statistics are provided in Table 1. For more detailed information we refer the interested readers to the original sources [25,57].

Table 1. Statistical specifics of the benchmark datasets used to evaluate ADD. Both datasets have significantly fewer notable visual artifacts than other DeepFake datasets [25,57].

Dataset	Class	Videos	Frames	Source	Train	Test	Val.
Celeb-DF (V2)	Pristine	590	225.4 K	YouTube	632	62	196
	DeepFake	5639	2116.8 K	DF	4736	536	340
WildDeepFake	Pristine	3805	680 K	Internet	3044	380	381
	DeepFake	3509	500 K	Internet	2807	350	351

4.2. Baseline Architectures

In the following, we briefly review four state-of-the-art deep learning models used in this study. These models are building the backbone of different configurations of ADD.

VGG19 Structure. The Visual Geometry Group (VGG) network is a type of deep convolutional neural network comprising 19 layers structured starting with five blocks of convolutional layers followed by three fully connected layers. Each convolutional layer contains a 3×3 kernel with a stride of 1 and padding of 1 to maintain the input–output dimensional match. Each of these convolutional layers are followed by a rectified linear unit (ReLU) activation and a max-pooling operation to reduce the spatial dimension. Max pooling layers employ a 2×2 kernel with a stride of 2 and no padding to reduce the size by 50%. Afterward, two fully connected layers with 4096 ReLU activated units are used before the final fully connected softmax classifier layer [64].

ResNet Structure. The Residual Networks (ResNets) [65] are a type of deep convolutional neural network where blocks of convolutional layers are skipped using shortcut connections. In this architecture, the down-sampling process takes place at convolutional layers with a stride of 2, after which batch normalization is performed. Finally, a ReLU activation is applied. The architecture has 101 layers in total, where the network ends with a fully connected layer with softmax activation [65].

Xception Structure. Xception is a convolutional neural network based on separable convolutions with residual connections. This model is composed of 71 deep layers, with an image input size of 299 by 299.

MobileNet Structure. MobileNet is a lightweight deep learning model developed using a depth-wise separable convolution architecture [66]. MobileNet architecture comprises 19 bottleneck layers consisting of three convolution operations, including 1×1 convolution, 3×3 depth-wise convolution, and 1×1 point-wise convolution. While the 1×1 convolution enriches the features through increasing number of channels, the 3×3 depth-wise convolution reduces computing costs by separating the feature filtering process. The separated features are then combined at point-wise convolution [66].

4.3. Implementation Specifics

Here, the implementation and characteristics of the ADD for reproducibility purposes are provided.

Implementation. All baseline models along with various configurations of ADD are implemented using the *PyTorch* machine learning library and trained using Stochastic Gradient Descent *SGD* optimizer [67] with a learning rate of 10^{-3}, momentum of 0.9, weight decay of 10^{-5}, and epoch number of 20 to minimize the softmax-cross-entropy loss. Moreover, we used mini-batch approaches with different mini-batch sizes for different deep network training process models. Mini-batch sizes are ranging from 8 for XceptionNet to 64 for VGG architecture on 4 NVIDIA Titan-V Graphics Processing Units (GPUs).

Experimental Setup. All experiments were conducted on two Lambda Quad deep learning workstations. Each workstation was equipped with Ubuntu 18.04 OS, Intel Xeon E5-1650 v4 CPU, 64 GB DDR4 RAM, 2TB SSD, 4TB HDD, and 4 NVIDIA Titan-V Graphics Processing Units (GPUs).

4.4. Evaluation Metrics

Performance of the ADD was evaluated against three different evaluation metrics, namely accuracy rate, recall, and area under the Receiver Operation Characteristic curve (ROC-AUC) at the frame level for all key frames. Although accuracy rate is easy to interpret, it might not provide a good insight for highly imbalanced datasets. Therefore, ROC-AUC metric was utilized to demonstrate how well the detection model performed on both DeepFake and pristine data distributions. Furthermore, recall metric was employed to reflect how well the model predicts manipulated videos, as missing a fake video is a costly mistake with potentially further adverse impacts. Additionally, all trained models will be published upon the acceptance of the paper. Having acquired the three metrics for ranking baseline models, they are ranked in three different manners and compared to the ground truth ranking attained from ADD on target test set. For fair comparison, all models were trained on the same training data and tested on the same hold-out test set.

5. Results & Discussion

This section provides a detailed discussion on the performance of the proposed DeepFake detection method. The performance of the ADD is evaluated based on three different evaluation metrics including detection accuracy, ROC, and recall. In our analysis, we focus on the DeepFake detection task at the level of each frame; hence, all reported results in this study are based on frame-wise detection tasks. First, the obtained results from simulations with/without ADD framework on Celeb-DF (V2) and WildDeepFake detection tasks are discussed for each baseline model to highlight the impact of ADD. Second, the performance of the ADD is compared to state-of-the-art DeepFake detection techniques.

5.1. ADD's Impact

To better understand the impact of the presented framework, ADD, we compared the performance of each baseline architecture with and without ADD using Celeb-DF (V2) and WildDeepFake benchmark datasets on frame-wise DeepFake detection problems, as reported in Table 2.

Table 2. The performance of the proposed ADD framework on Celeb-DF (V2) and WildDeepFake DeepFake detection benchmarks using four different baseline architectures.

Backbone	Method	Celeb-DF (V2)			WildDeepFake		
		ACC	ROC	Recall	ACC	ROC	Recall
ResNet	Baseline	88.47	91.47	84.76	58.73	59.12	58.81
	ADD	98.37	98.65	97.59	78.15	78.01	78.24
Xception	Baseline	94.54	95.42	92.69	69.25	69.78	69.01
	ADD	97.32	97.84	96.03	80.13	80.31	79.93
VGG	Baseline	95.53	96.71	93.93	60.92	61.12	60.71
	ADD	97.93	98.01	97.35	77.83	77.54	77.61
MobileNet	Baseline	91.72	91.68	91.35	61.78	61.54	61.23
	ADD	94.63	94.76	93.89	78.67	78.31	78.15

Celeb-DF (V2). The obtained results for conducted simulations using Celeb-DF (V2) dataset for each model architecture are shown in Figure 5. As it can be observed, baseline models did not perform well; the best baseline model reached 95.53% detection accuracy at best, via VGG structure, which is not acceptable in the DeepFake detection task. While

performances of the baseline models were poor, their performances were boosted by considerably large margins once upgraded to the ADD framework. For example, the performance of the vanilla detection model with ResNet architecture improved from 88.47% detection accuracy rate to 98.37% on the same model with the ADD framework, which is around a 10% improvement on the detection rate. Obtained results from the experiments clearly demonstrate the outstanding impact of the proposed attention-based framework in this study for enhanced DeepFake detection.

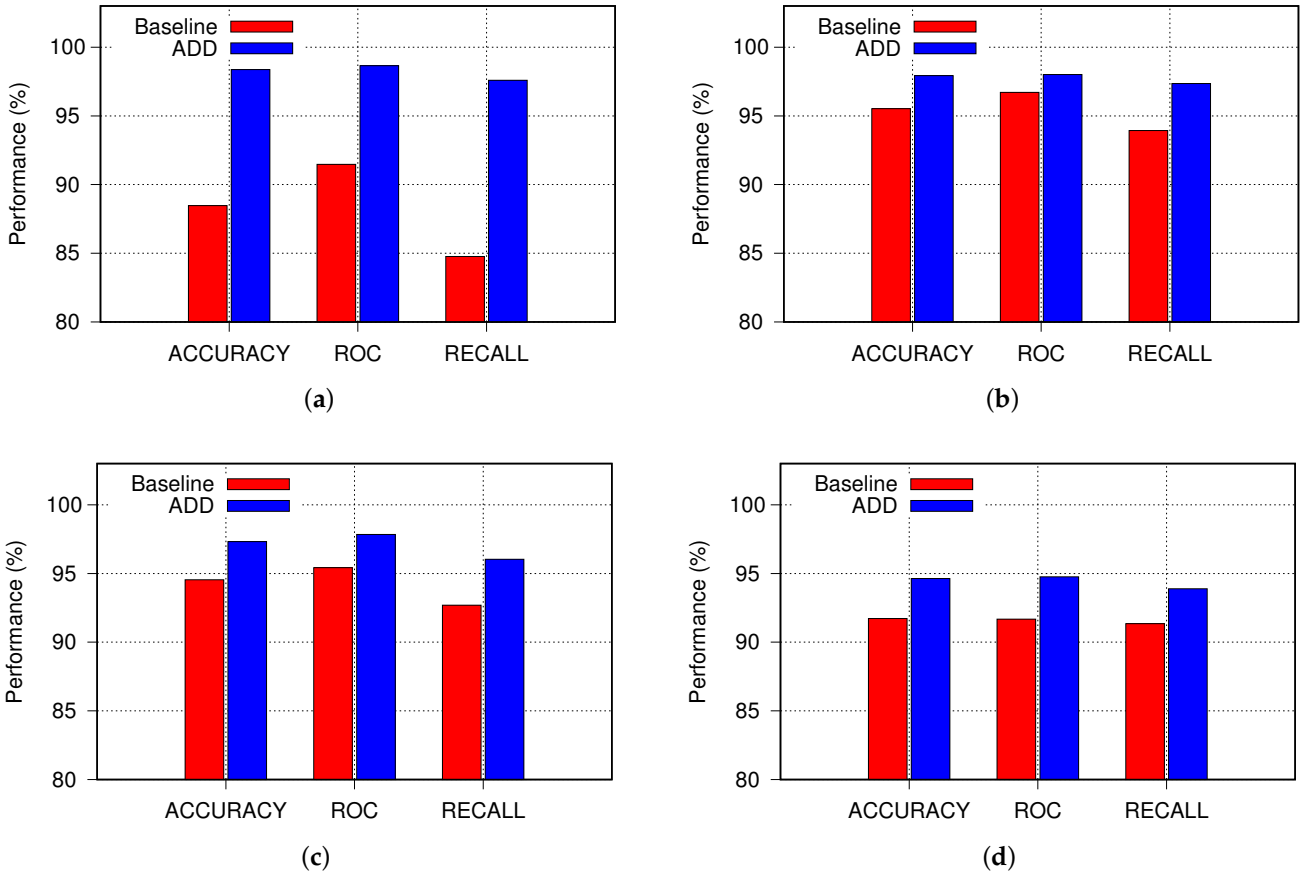

Figure 5. The DeepFake detection method results on Celeb-DF (V2) using four different baseline models and ADD configurations. As shown, all four configurations of ADD can significantly improve all three evaluation metrics, i.e., accuracy rate, ROC, and recall. Note that all models are trained and evaluated against the same datasets. (**a**) Backbone: ResNet Model; (**b**) Backbone: Xception Model; (**c**) Backbone: VGG Model; (**d**) Backbone: MobileNet Model.

WildDeepFake. WildDeepFake dataset is more challenging to be detected compared to virtual DeepFake; therefore, the effectiveness of detectors developed on virtual DeepFake datasets can be limited when applied to wild DeepFake. A similar set of experiments are conducted using WildDeepFake to evaluate the performance of the proposed method on a more challenging DeepFake detection task. The obtained results from these experiments are illustrated in Figure 6. A similar pattern to previous experiments was observed, which confirms the effectiveness of ADD framework on improving the detection performance of all four baseline models. As it can be seen, vanilla models that were not equipped with an attention mechanism did not offer an acceptable detection accuracy rate, not more than 69%, which is extremely low in the DeepFake detection field. However, all configurations of ADD were able to improve the evaluation metrics by significantly large margins. For example, ADD with Xception baseline architecture detected DeepFake with 79.23% detection accuracy. This result is outstanding compared to existing state-of-the-art DeepFake detection methods.

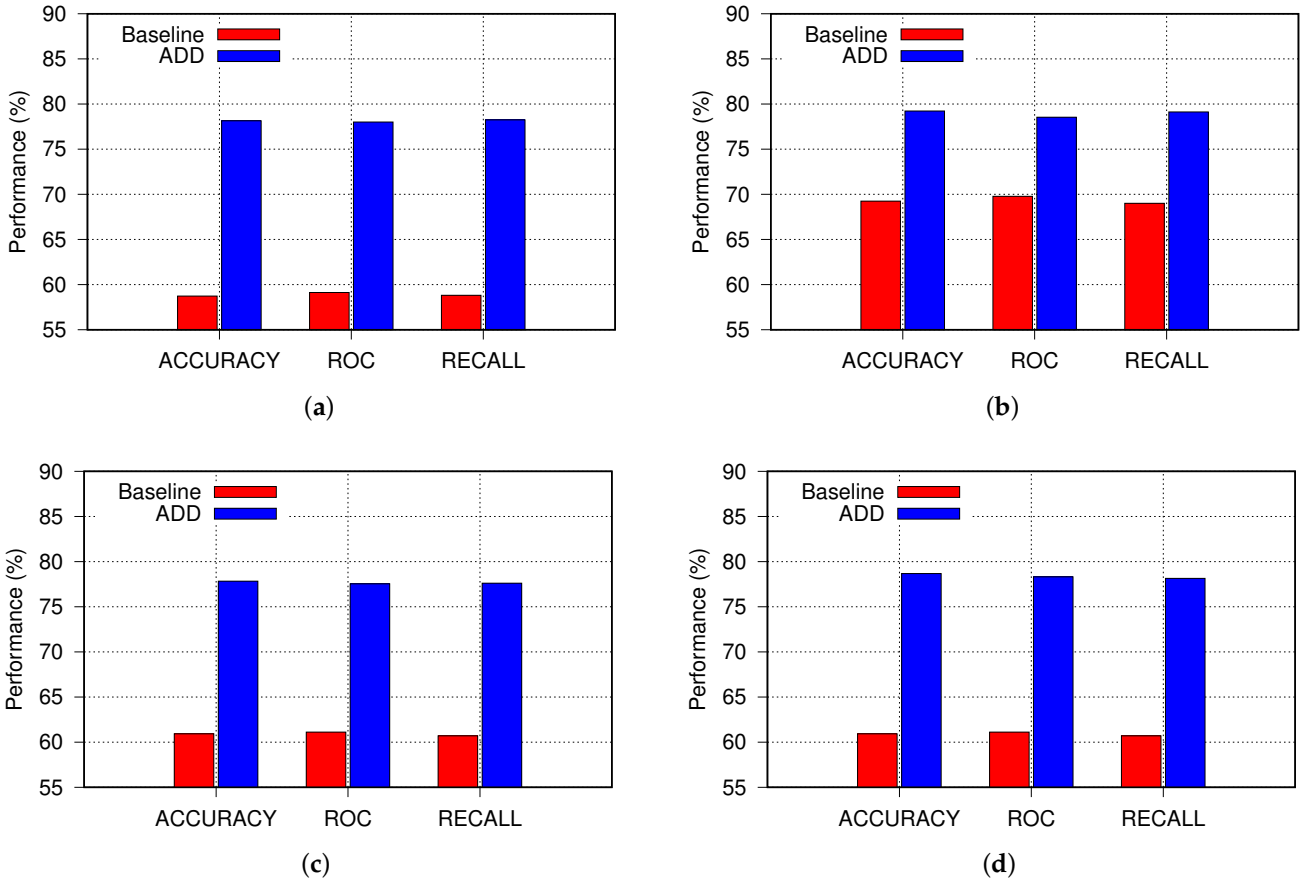

Figure 6. The results of the DeepFake detection method on WildDeepFake using four different baseline models and ADD configurations. As shown, all four configurations of ADD can improve evaluation metrics by a significantly outstanding margin. (**a**) Backbone: ResNet Model; (**b**) Backbone: Xception Model; (**c**) Backbone: VGG Model; (**d**) Backbone: MobileNet Model.

5.2. Comparison with State-of-the-Art Methods

This section is devoted to comparing the performance of the ADD against state-of-the-art methods on DeepFake detection tasks. While we have reported different evaluation metrics in our analysis, we follow the reported metrics in the literature for comparison. The obtained results on Celeb-DF (V2) are reported based on the AUC score in the literature; therefore, we compare the AUC score of ADD with that of the literature, as shown in Table 3. It can be observed that all configurations of ADD outperformed the state-of-art detection AUC score with significantly large margins. In particular, ADD with ResNet baseline architecture achieved an AUC score of 98.65%, which is more than a 7% improvement compared to FakeCatcher [44].

Since detection accuracy rate is the only reported evaluation metric regarding the performance of detection techniques on WildDeepFake benchmark [57], we used the same metric for our comparison study. The obtained results from our experiments along with other approaches on this particular dataset are reported in Table 4. It can be observed that while most of the previous studies were bound to below 70% accuracy rates, our proposed ADD framework boosted the performance of all baseline models above 77%. For instance, a configuration of ADD with Xception architecture was able to further improve the DeepFake detection performance on WildDeepFake dataset and achieve 80.13%. Overall, it can be observed that all configurations of ADD outperformed the state-of-art DeepFake detection methods with significantly large margins.

Table 3. Comparing the performance of ADD against state-of-the-art DeepFake detection models on Celeb-DF (V2). Note that reported results in rows 1–8 are from [25].

Models	AUC (%)
Two-stream [68]	53.8
Meso4 [33]	54.8
HeadPose [61]	54.6
FWA [11]	56.9
VA-MLP [45]	55.0
Xception-c40 [58]	65.5
Multi-task [47]	54.3
Capsule [69]	57.5
TBRN [70]	73.41
Face X-ray [71]	80.58
PPA [72]	83.10
FakeCatcher [44]	91.50
ADD-ResNet (ours)	**98.37**
ADD-Xception (ours)	97.32
ADD-VGG (ours)	97.93
ADD-MobileNet (ours)	94.63

Table 4. Comparing the detection accuracy rate of ADD against state-of-the-art DeepFake detection models on WildDeepFake dataset. Note that reported metrics in rows 1–8 are from [57].

Models	ACC (%)
AlexNet [73]	60.37
VGG16 [74]	60.92
ResNetV2-50 [75]	63.99
ResNetV2-101 [75]	58.73
ResNetV2-152 [75]	59.33
Inception-v2 [76]	62.12
MesoNet-1 [33]	60.51
MesoNet-4 [33]	64.47
MesoNet-inception [33]	66.03
XceptionNet [77]	69.25
ADDNet-2D [57]	76.25
ADDNet-3D [57]	65.50
ADD-ResNet (ours)	78.15
ADD-Xception (ours)	**79.23**
ADD-VGG (ours)	77.83
ADD-MobileNet (ours)	78.67

6. Conclusions

This paper presents a DeepFake detection method, ADD, that exploits the fine-grained and spatial locality attributes of the AI-synthesized videos to boost detection performance. Potentially manipulated areas of the input image and corresponding features are first extracted, and then the detection model is forced to focus more on those manipulated regions for decision making. ADD performs this task by imposing extra supervision on instance interpretation in the learning procedure. The performance of ADD is evaluated against two recently introduced challenging datasets for DeepFake forensics, i.e., Celeb-DF (V2) and WildDeepFake. For example, ADD with ResNet architecture is able to detect DeepFakes with more than 98.3% AUC on Celeb-DF (V2), outperforming state-of-the-art DeepFake detection methods.

Author Contributions: A.K. came up with the idea, ran the experiments, and wrote the manuscript. J.-S.Y. provided technical feedback and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by Florida Center for Cybersecurity.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
2. Antipov, G.; Baccouche, M.; Dugelay, J.L. Face aging with conditional generative adversarial networks. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 2089–2093.
3. Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2387–2395.
4. Vaccari, C.; Chadwick, A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Soc. Media+ Soc.* **2020**, *6*. [CrossRef]
5. Toews, R. Deepfakes Are Going to Wreak Havoc on Society. We Are not Prepared. 2020. Available online: https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared (accessed on 15 January 2021).
6. Turton, W.; Martin, A. How Deepfakes Make Disinformation More Real Than Ever. *Bloomberg* **2020**, 199–217. Available online: https://www.bloombergquint.com/technology/how-deepfakes-make-disinformation-more-real-than-ever-quicktake (accessed on 15 January 2021)
7. Ingram, D. A Face-Swapping App Takes off in China, Making Ai-Powered Deepfakes for Everyone. 2019. Available online: https://www.wautom.com/2019/09/zao-a-face-swapping-app-takes-off-in-china-making-ai-powered-deepfakes-for-everyone/ (accessed on 15 January 2021).
8. Greengard, S. Will deepfakes do deep damage? *Commun. ACM* **2019**, *63*, 17–19. [CrossRef]
9. Korshunov, P.; Marcel, S. *Deepfake Detection: Humans vs. Machines*; The European Association for Biometrics (EAB): Bussum, The Netherlands, 2020.
10. Guarnera, L.; Giudice, O.; Battiato, S. DeepFake Detection by Analyzing Convolutional Traces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 666–667.
11. Li, Y.; Lyu, S. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 46–52.
12. Maras, M.H.; Alexandrou, A. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *Int. J. Evid. Proof* **2019**, *23*, 255–262. [CrossRef]
13. Day, C. The future of misinformation. *IEEE Ann. Hist. Comput.* **2019**, *21*, 108. [CrossRef]
14. Fletcher, J. Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance. *Theatre J.* **2018**, *70*, 455–471. [CrossRef]
15. Chesney, B.; Citron, D. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.* **2019**, *107*, 1753. [CrossRef]
16. Thies, J.; Zollhöfer, M.; Nießner, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–12. [CrossRef]
17. Nirkin, Y.; Masi, I.; Tuân, A.T.; Hassner, T.; Medioni, G. On face segmentation, face swapping, and face perception. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 98–105.
18. Masi, I.; Trân, A.T.; Hassner, T.; Leksut, J.T.; Medioni, G. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 579–596.
19. Masi, I.; Trân, A.T.; Hassner, T.; Sahin, G.; Medioni, G. Face-specific data augmentation for unconstrained face recognition. *Int. J. Comput. Vis.* **2019**, *127*, 642–667. [CrossRef]
20. Suwajanakorn, S.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph. (TOG)* **2017**, *36*, 1–13. [CrossRef]
21. Pumarola, A.; Agudo, A.; Martinez, A.M.; Sanfeliu, A.; Moreno-Noguer, F. Ganimation: Anatomically-aware facial animation from a single image. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 818–833.
22. Nirkin, Y.; Keller, Y.; Hassner, T. FSGAN: Subject agnostic face swapping and reenactment. In Proceedings of the IEEE International Conference on COMPUTER Vision, Seoul, Korea, 27–28 October 2019; pp. 7184–7193.
23. Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; Theobalt, C. Deep video portraits. *ACM Trans. Graph. (TOG)* **2018**, *37*, 1–14. [CrossRef]

24. Natsume, R.; Yatagawa, T.; Morishima, S. Fsnet: An identity-aware generative model for image-based face swapping. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 117–132.
25. Li, Y.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In Proceedings of the IEEE Conference on Computer Vision and Patten Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
26. Du, M.; Pentyala, S.; Li, Y.; Hu, X. Towards Generalizable Deepfake Detection with Locality-aware AutoEncoder. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Gold Coast, Australia, 1–5 November 2020; pp. 325–334.
27. Yadav, D.; Salmani, S. Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network. In Proceedings of the 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 15–17 May 2019; pp. 852–857.
28. Shelke, N.A.; Kasana, S.S. A comprehensive survey on passive techniques for digital video forgery detection. *Multimed. Tools Appl.* **2021**, *80*, 6247–6310. [CrossRef]
29. Mirsky, Y.; Lee, W. The creation and detection of deepfakes: A survey. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–41. [CrossRef]
30. Li, Y.; Chang, M.C.; Lyu, S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7.
31. Huh, M.; Liu, A.; Owens, A.; Efros, A.A. Fighting fake news: Image splice detection via learned self-consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 101–117.
32. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2307–2311.
33. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7.
34. Faceswap. Faceswap: Deepfakes Software for All. Available online: https://faceswap.dev/ (accessed on 15 January 2021).
35. FakeApp. FakeApp 2.2.0—Download for PC Free. 2019. Available online: https://www.malavida.com/en/soft/fakeapp/ (accessed on 15 January 2021).
36. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
37. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, Sydney, Australia, 17 July 2017; pp. 1857–1865.
38. Lu, Y.; Tai, Y.W.; Tang, C.K. Attribute-guided face generation using conditional cyclegan. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 282–297.
39. Shaoan, L. Faceswap-GAN. 2020. Available online: https://github.com/shaoanlu/faceswap-GAN (accessed on 15 January 2021).
40. Wang, T.C.; Liu, M.Y.; Tao, A.; Liu, G.; Kautz, J.; Catanzaro, B. Few-shot video-to-video synthesis. In Proceedings of the Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
41. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Liu, G.; Tao, A.; Kautz, J.; Catanzaro, B. Video-to-Video Synthesis. *Adv. Neural Inf. Process. Syst.* **2018**, 1144–1156.
42. Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; Sebe, N. First order motion model for image animation. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 7137–7147.
43. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv* **2019**, arXiv:1912.13457.
44. Ciftci, U.A.; Demir, I.; Yin, L. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef]
45. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 7–11 January 2019; pp. 83–92.
46. Cozzolino, D.; Thies, J.; Rössler, A.; Riess, C.; Nießner, M.; Verdoliva, L. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv* **2018**, arXiv:1812.02510.
47. Nguyen, H.H.; Fang, F.; Yamagishi, J.; Echizen, I. Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos. In Proceedings of the 2019 IEEE 10th International Conference on Biometrics: Theory, Applications and Systems (BTAS), Tampa, FL, USA, 23–26 September 2019; pp. 1–8. [CrossRef]
48. Rana, M.S.; Sung, A.H. Deepfakestack: A deep ensemble-based learning technique for deepfake detection. In Proceedings of the 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), New York, NY, USA, 1–3 August 2020; pp. 70–75.
49. Güera, D.; Delp, E.J. Deepfake video detection using recurrent neural networks. In Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November 2018; pp. 1–6.
50. Sabir, E.; Cheng, J.; Jaiswal, A.; AbdAlmageed, W.; Masi, I.; Natarajan, P. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)* **2019**, *3*, 80–87.
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 770–778.
52. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
53. Yu, N.; Davis, L.S.; Fritz, M. Attributing fake images to gans: Learning and analyzing gan fingerprints. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 7556–7566.

54. Đorđević, M.; Milivojević, M.; Gavrovska, A. DeepFake Video Analysis using SIFT Features. In Proceedings of the 2019 27th Telecommunications Forum (TELFOR), Belgrade, Serbia, 26–27 November 2019; pp. 1–4.
55. Kaur, S.; Kumar, P.; Kumaraguru, P. Deepfakes: Temporal sequential analysis to detect face-swapped video clips using convolutional long short-term memory. *J. Electron. Imaging* **2020**, *29*, 033013. [CrossRef]
56. Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues. In Proceedings of the 28th ACM International Conference on Multimedia, 2020; pp. 2823–2832.
57. Zi, B.; Chang, M.; Chen, J.; Ma, X.; Jiang, Y.G. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In Proceedings of the 28th ACM International Conference on Multimedia, 2020; pp. 2382–2390.
58. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 1–11.
59. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5203–5212.
60. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874.
61. Yang, X.; Li, Y.; Lyu, S. Exposing deep fakes using inconsistent head poses. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8261–8265.
62. Korshunov, P.; Marcel, S. Deepfakes: A new threat to face recognition? assessment and detection. *arXiv* **2018**, arXiv:1812.08685.
63. Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Ferrer, C.C. The deepfake detection challenge (dfdc) preview dataset. *arXiv* **2019**, arXiv:1910.08854.
64. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
65. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
66. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
67. Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the importance of initialization and momentum in deep learning. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1139–1147.
68. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Two-stream neural networks for tampered face detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1831–1839.
69. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Use of a capsule network to detect fake images and videos. *arXiv* **2019**, arXiv:1910.12467.
70. Masi, I.; Killekar, A.; Mascarenhas, R.M.; Gurudatt, S.P.; AbdAlmageed, W. Two-Branch Recurrent Network for Isolating Deepfakes in Videos. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 667–684.
71. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face X-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5001–5010.
72. Charitidis, P.; Kordopatis-Zilos, G.; Papadopoulos, S.; Kompatsiaris, I. Investigating the Impact of Pre-processing and Prediction Aggregation on the DeepFake Detection Task. In Proceedings of the Truth and Trust Conference, Virtual, 15–17 October 2020.
73. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
74. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In the Proceedings of the International Conference on Learning Representation, San Diego, CA, USA, 7–9 May 2015.
75. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 630–645.
76. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
77. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.