applied sciences

Article

# DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer

**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:** The ever-growing threat of deepfakes and large-scale societal implications has propelled the development of deepfake forensics to ascertain the trustworthiness of digital media. A common theme of existing detection methods is using Convolutional Neural Networks (CNNs) as a backbone. While CNNs have demonstrated decent performance on learning local discriminative information, they fail to learn relative spatial features and lose important information due to constrained receptive fields. Motivated by the aforementioned challenges, this work presents DFDT, an end-to-end deepfake detection framework that leverages the unique characteristics of transformer models, for learning hidden traces of perturbations from both local image features and global relationship of pixels at different forgery scales. DFDT is specifically designed for deepfake detection tasks consisting of four main components: patch extraction & embedding, multi-stream transformer block, attention-based patch selection followed by a multi-scale classifier. DFDT's transformer layer benefits from a re-attention mechanism instead of a traditional multi-head self-attention layer. To evaluate the performance of DFDT, a comprehensive set of experiments are conducted on several deepfake forensics benchmarks. Obtained results demonstrated the surpassing detection rate of DFDT, achieving 99.41%, 99.31%, and 81.35% on FaceForensics++, Celeb-DF (V2), and WildDeepfake, respectively. Moreover, DFDT's excellent cross-dataset & cross-manipulation generalization provides additional strong evidence on its effectiveness.

**Keywords:** cybersecurity; deep learning; deepfake detection; vision transformer

check for updates

**Citation:** Khormali, A.; Yuan, J.-S. DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer. *Appl. Sci.* **2022**, *12*, 2953. https://doi.org/10.3390/app12062953

Academic Editor: Alberto Bartoli

Received: 8 February 2022
Accepted: 5 March 2022
Published: 14 March 2022

## 1. Introduction

The recent advances in the field of Artificial Intelligence (AI), particularly Generative Adversarial Networks (GANs) [1,2] and the abundance of training samples, along with robust computational resources [3], have significantly propelled the field of AI-generated fake information in all kinds, e.g., deepfakes. Deepfakes are synthesized yet super-realistic images and videos generated through combining, merging, superimposing, or replacing the facial area of images/videos leveraging advanced techniques from computer vision and deep learning domains [4]. Deepfakes are among the most sinister types of misinformation, posing large-scale and severe security and privacy risks targeting critical governmental institutions and ordinary people across the world [5,6]. Furthermore, deepfake generation algorithms are constantly evolving and have become a bullet point for adversarial entities to perpetuate and disseminate criminal content in various forms, including ransomware, digital kidnapping, etc. [7].

The ever-growing threat of deepfakes and large-scale societal implications have driven the development of deepfake forensics to ascertain the trustworthiness and authenticity of digital media. Different deepfake detection approaches have been proposed to address this challenge [8,9]. Early deepfake detection algorithms were primarily based on hand-crafted features, and visible artifacts, such as inconsistency in head poses [10], eye blinking [11] and face wrapping artifacts [12]. However, as deepfakes become more deceitful and sophisticated, deepfake detection algorithms are advancing. The fact that deepfakes are

GAN-generated digital content and not actual events captured by a camera implies that they still can be detected using advanced AI models [13]. Furthermore, it has been proven that deep neural networks tend to achieve better performance than traditional image forensic tools [9]. Typical components of most state-of-the-art deepfake detection approaches are convolutional neural networks, and facial regions cropped out of an entire image [14–16]. Unique characteristics of the convolutional operator in deep CNNs have enabled them to demonstrate strong capabilities on capturing minor visual artifacts, yielding decent detection results [7,17,18]. Although CNNs have proven themselves solid candidates for learning local information of the image, they still miss capturing pixels' spatial interdependence due to constrained receptive fields.

Almost all deepfakes are created by applying alterations to the facial area while leaving other regions intact. Therefore, in addition to local image features, every pixel's global relationship provides essential information regarding the intensity and extension of manipulations. This information could be augmented to boost the performance of the deepfake detection algorithm and bring better insight into the location of the forgeries. At the same time, different deepfake generation techniques target different proportions and regions of the facial area to be manipulated, ranging from small regions such as color mismatch in lips to larger areas that extend throughout the image like face boundaries in face-swapping approaches. Therefore, it is vital to successfully identify forged pixels to extract better discriminative features in a scalable manner. Motivated by the aforementioned challenges, a multi-stream deepfake detection framework is presented that incorporates pixels' spatial interdependence in a global context with local image features in a scalable scheme using unique characteristics of transformer models on learning global relationship of pixels. Transformer models have proven their strong capabilities on learning long-term dependency on natural language processing tasks [19–21], and more recently on computer vision tasks [22–24].

**Objectives.** Although the deepfake detection task has gained massive attention within the last couple of years, the mainstream detection methods rely on localized features and CNN-based structures. Surprisingly only a few research works have been conducted on the intersection of vision transformers and facial forgery detection. The main goal of this study is to present a digital media authentication system leveraging the unique characteristics of vision transformers on modeling the global relationship of pixels in different manipulation scales. While CNNs fail to learn relative spatial information and lose essential data in pooling layers, vision transformers' global attention mechanism enables the network to learn higher-level information much faster, which leads to more promising performances in less computational time. Furthermore, in digital media forensics, it is of vital importance not only to detect deepfakes but also it is equally important to recognize specific parts of the image that has been forged. This goal can be achieved by extracting hidden traces and intrinsic representations from the image's manipulated regions.

**Contributions.** While existing deepfake detection approaches are primarily dependent on CNN-based structures, this work presents an end-to-end deepfake detection framework leveraging the unique characteristics of transformer models. DFDT discovers hidden traces of perturbations from both local image features and global relationship of pixels at different forgery scales. Unlike previous studies that are limited to either a direct application of vision transformers or still heavily rely on CNN-based models as backbone [25,26], this work presents a transformer model that was mainly developed for the deepfake detection task. A comprehensive set of analyses are conducted to assess the performance of the proposed method from various perspectives, including intra-dataset performance, cross-dataset & cross-manipulation generalization, and various ablation studies. The key contributions of this work are summarized as follows:

- An end-to-end deepfake detection framework, DFDT, is developed leveraging the unique characteristics of transformer models on learning hidden traces of perturbations from both local image features and global relationship of pixels at different forgery scales.

- DFDT is designed explicitly for deepfake detection tasks. DFDT comprises four main components, including patch extraction & embedding, multi-stream transformer block, attention-based patch selection followed by a multi-scale classifier. DFDT's transformer layer benefits from the re-attention mechanism instead of the traditional multi-head self-attention layer.
- A comprehensive set of experiments are conducted on seven deepfake forensics benchmarks to evaluate the performance of the DFDT. Experimental results demonstrated the surpassing detection rate of the DFDT, achieving 99.41%, 99.31%, and 81.35% on FaceForensics++, Celeb-DF (V2), and WildDeepfake, respectively. Moreover, DFDT's excellent cross-dataset & cross-manipulation generalization provides additional strong evidence on its effectiveness.

**Organization.** The rest of the paper is organized as follows. Section 2 provides a brief discussion on recent significant works on deepfake generation and detection techniques. The outline of the presented approach, DFDT, along with its main components patch extraction & embedding, multi-stream transformer block, and attention-based patch selection are presented in Section 3. Overall evaluation settings, including datasets, implementation specifics, and evaluation metrics are described in Section 4. The obtained experimental results on DFDT are discussed and compared to its counterparts in Section 5. Finally, concluding remarks are drawn in Section 6.

## 2. Related Work

A brief description of recent advancements in deepfake analysis domain is provided here.

**DeepFake Generation.** Although early deepfake generation techniques were mostly based on traditional vision and voice impersonation methods [3,27,28], most recent techniques benefit from the unique generation capabilities of GANs. For instance, Zhu et al. [29], and Kim et al. [30] utilized cycle-consistent GANs to generate deepfakes such that it maintains the facial expressions of the target while swapping the identities of source and target. Furthermore, Lu et al. [31] presented identity-guided conditional CycleGAN to convert low-resolution facial images to high-resolution images. Similarly, Kim et al. [32] introduced a deep video portraits method was introduced to transfer both facial expression and 3D poses of the source image into the target image. Moreover, Li et al. [33] presented a high-quality face replacement approach through FaceShifter that exploits a learning method based on heuristic error acknowledging refinement network. As the main scope of this study is on deepfake detection techniques, more interested readers are referred to [8,9] for more detailed information on state-of-the-art deepfake generation techniques.

**Deepfake Detection.** While the deepfake detection task has been studied from different perspectives, this study mainly explores the AI-driven deepfake detection approaches. Given the importance and huge threat of deepfake technologies, a large body of work is devoted to devising high-performance and resilient detection technologies. While early-stage detection techniques mainly focused on handcrafted features, i.e., blinking inconsistencies [12], biological signals [15], and unrealistic details [34], more recent techniques are developed using advanced deep learning networks. For example, Afchar et al. [35] introduced *MesoNet* as a deepfake detection algorithm that is composed of a shallow convolutional network and intermediate level of features. A detection approach based on auto-encoder architecture and transfer learning, *forensictransfer*, was presented by Cozzolino et al. [36]. Similarly, *Capsule-Forensics* architecture was introduced by Nguyen et al. [16] for better detection of AI-generated images and videos. Furthermore, an ensemble learning approach was employed to improve deepfake detection composite model by Rana et al. [37]. Additionally, Kaur et al. [38] proposed a detection approach based on sequential temporal analysis and convolutional long short-term memory networks. Wang et al. [39] evaluated the cross-dataset generalization capability of their detection model, which was trained over ProGAN and tested on other datasets. Mittal et al. [40] proposed a multi-modal approach composed of audio and video modalities to tackle deepfake detection tasks. Furthermore, Jian et al. [14] introduced a hierarchical classification approach that can recognize deepfakes at three different levels, including manipulated images, retouched from GAN-based images, and specific GAN architecture. Other researchers have investigated deepfake detection tasks from a fine-grained visual classification point of view, specifically attention-based techniques. For instance, Du et al. [13] proposed a deepfake detection method from a fine-grained visual classification angle that is built using an auto-encoder architecture. Furthermore, Khormali and Yuan [18] have presented an attention-based deepfake detection approach utilizing two different modules, i.e., Face close-up and Face Shut-off, to force the model to extract more discriminative information from other parts of the facial region. Quan et al. [41] presented a progressive transfer learning algorithm to tackle face spoofing attacks using only a limited number of training samples. The presented face anti-spoofing method benefits from a temporal consistency constraint to verify the reliability of pseudo labels of selected data. While a large body of work is focused on CNN-based approaches for deepfake detection, only minimal effort has been devoted to investigating more advanced technologies such as vision transformers for deepfake detection.

**Vision Transformer.** Transformer networks were primarily designed to learn long-range contextual information to solve natural language processing tasks, e.g., text classification, machine translation. Transformers are extremely scalable and have demonstrated remarkable performances on learning dependency among frames of large-scale datasets, e.g., BERT [20], BioBERT [21], and GPT-3 [42]. Inspired by the great performance of transformers in NLP tasks, they have been recently extend to computer vision and multi-modal vision-language tasks, such as image classification [24,43], object detection [23,44], and image segmentation [45]. On the other hand, minimal effort is devoted to exploring vision transformers for deepfake detection. Existing methods still highly depend on CNNs for feature extraction. The community lacks an end-to-end vision-transformer framework designed explicitly for deepfake detection tasks. For example, Khan and Dai [46] presented a video transformer with an incremental learning approach for deepfake detection. Their design benefits from XceptionNet [47] as a backbone for image feature extraction and 12 transformer blocks for feature learning. Similarly, Wodajo and Atnafu [26] presented a convolutional vision transformer that uses CNNs as a feature extractor and a transformer block as a classifier. Furthermore, Heo et al. [25] proposed an scheme based on vision transformer and distillation that is build based on EfficientNet [48] features. Therefore, to fill this research gap in the domain, an end-to-end transformer-based framework is developed explicitly and developed for the deepfake detection problem in this study.

## 3. Methodology

A detailed description of the building blocks of the proposed multi-stream transformer-based deepfake detection framework, DFDT, is presented in this section. DFDT consists of four main components, including patch & embeddings extraction (Section 3.1), a multi-stream transformer block (Section 3.3), attention-based patch selection (Section 3.3), followed by a multi-scale classifier. The overall framework of the DFDT is depicted in Figure 1.

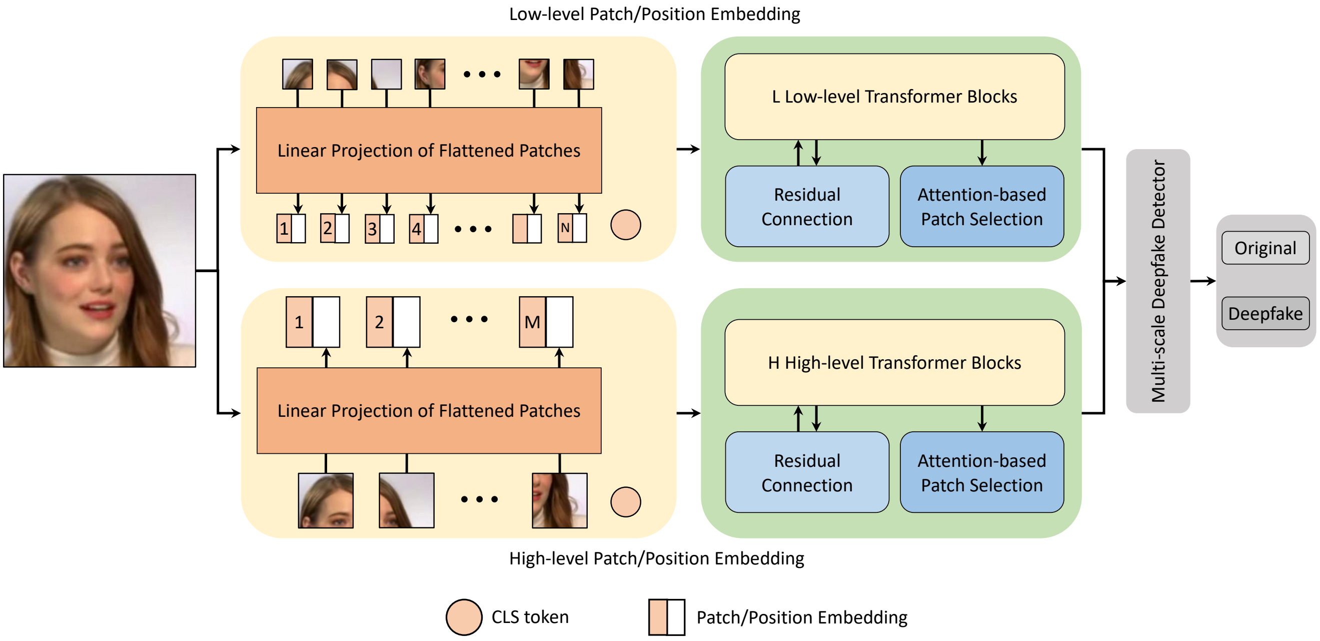

Low-level Patch/Position Embedding
Linear Projection of Flattened Patches
1
2
3
4
N
L Low-level Transformer Blocks
Residual Connection
Attention-based Patch Selection
1
2
M
Linear Projection of Flattened Patches
H High-level Transformer Blocks
Residual Connection
Attention-based Patch Selection
High-level Patch/Position Embedding
Multi-scale Deepfake Detector
Original
Deepfake
CLS token
Patch/Position Embedding

**Figure 1.** An overview of the proposed end-to-end transformer-based deepfake detection framework. DFDT is developed leveraging the unique characteristics of transformer models on learning hidden traces of perturbations from local image features and the global relationship of pixels at different forgery scales. Patch extraction & embedding, a multi-stream transformer block, attention-based patch selection, followed by a multi-scale classifier are the main components of the DFDT.

### *3.1. Patch Extraction & Embedding*

All mandatory preprocessing steps, including face localization, patch extraction, and patch/positional embeddings, are described in this section.

**Preprocessing & Face Localization.** In general image manipulation tasks, the forgeries might be scattered across regions of the image, however, in deepfake generation techniques manipulations are mainly limited to facial areas and the background area is left intact. Therefore, having trained a model based on only face area would not only reduce computational complexity but also would improve the model performance due to background noise reduction [49]. Therefore, first, 20% of frames are extracted for each input video in consecutive order, and then facial landmarks are extracted leveraging the RetinaFace [50], a state-of-the-art face detection method. Finally, the facial area on each frame is cropped, resized, and aligned using the calculated landmarks [51].

**Patch & Embedding Extraction.** Before analyzing images using transformer models, they need to be converted into a $D$-dimensional sequence of smaller patch embeddings. While non-overlapping patch extraction methods harm the neighboring local structures, the overlapping image splitting approach, where two adjacent patches share an area, helps preserve and learn the neighboring information of the local area much better. In other words, in the overlapping patch extraction approach, every given image with a resolution of $(H, W)$ and $C$ channels, $I \in \mathbb{R}^{H\times W\times C}$, is dissected into $N$ smaller image patches with a resolution of $(P, P)$ and $C$ channels, $I_p \in \mathbb{R}^{P\times P\times C}$, using a sliding window of stride $S$. Each adjacent patch shares an area of size $P \times (P - S)$.

$$N = N_H * N_W = \left\lfloor \frac{H - P + S}{S} \right\rfloor * \left\lfloor \frac{W - P + S}{S} \right\rfloor \tag{1}$$

The resulting patches are then flattened and projected into a latent $D$-dimensional embedding space $\mathbf{E} \in \mathbb{R}^{N\times D}$. Furthermore, to maintain positional information of each patch, patch embeddings are integrated with position embeddings $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$. The resulting sequence of embedded patches, **z**, serves as an effective input sequence for the transformer blocks [24].

$$\mathbf{z} = [I_{\text{cls}}\,; E] + E_{\text{pos}}, \quad E \in \mathbb{R}^{N\times D}, E_{\text{pos}} \in R^{(N+1)\times D} \tag{2}$$

### *3.2. Attention-Based Patch Selection*

Other researchers have demonstrated that transformer models cannot accurately represent the importance level of input tokens, especially in higher layers, due to a lack of token identifiability of the embeddings [52]. One solution to improve the transformer model's capability in capturing sensitive information is to pay more attention to discriminative patches within the training and inference phases, as depicted in Figure 2. In addition, attention-based mechanisms have demonstrated their strong capability in improving the performance of traditional CNN-based deepfake detection models [18]. Therefore, inspired by [53], a patch selection mechanism based on attention weights is l in this study. Generally, for a transformer model with $K$ attention heads and $L$ layers, the input feature to the last layer and attention weights of $l$th layer can be represented as (3) and (4), respectively.

$$\mathbf{Z}_{L-1} = \left[Z^0_{L-1}; Z^1_{L-1}, Z^2_{L-1}, \cdots, Z^N_{L-1}\right] \tag{3}$$

$$\begin{aligned} \mathbf{a}_l &= \left[a^0_l, a^1_l, a^2_l, \cdots, a^K_l\right] \qquad l \in 1, 2, \cdots, L-1 \\ \mathbf{a}^i_l &= \left[a^{i_0}_l; a^{i_1}_l, a^{i_2}_l, \cdots, a^{i_N}_l\right] \qquad i \in 0, 1, \cdots, K \end{aligned} \tag{4}$$

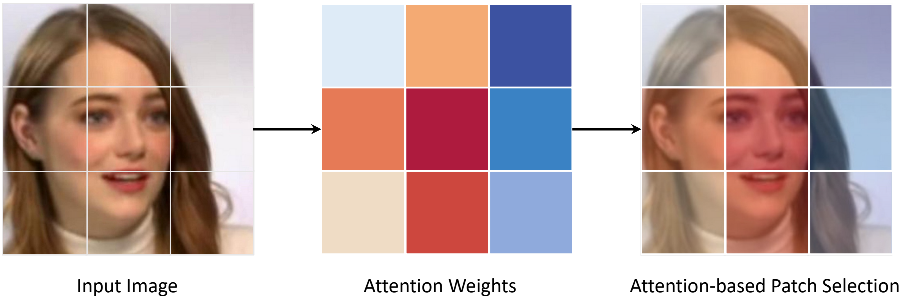


**Figure 2.** The attention-based patch selection module forces the transformer model to put more weight on key patches, while dismissing less informative regions.

It is critically important to learn how information propagates through different layers and ensure the correspondence of attention weights with input tokens. Therefore, the raw attention weights are fused throughout the transformer model using matrix multiplication as shown in (5). Note that $\mathbf{a}_{\text{final}}$ provides better discriminative information on selection of top patches compared to single-layer raw attention weights, i.e., $\mathbf{a}_l$ as in (4) [53]. Therefore, positional index of maximum attention weights corresponding to $K$ attention heads in $\mathbf{a}_{\text{final}}$ are calculated, i.e., $A_1, A_2, \cdots, A_K$ and utilized to extract corresponding key tokens in $\mathbf{z}_{L-1}$. Ultimately, the original input sequence in (3) is substituted with an updated sequence consisting of the concatenation of key tokens corresponding to more informative regions along with the classification token as (6). The new input sequence maintains the global information of the input image. This process makes sure that the model pays specific attention to the subtle traces of deepfakes while dismissing less discriminative regions like the background area.

$$\mathbf{a}_{\text{final}} = \prod_{l=0}^{L-1} \text{Softmax}\left(\mathbf{a}_l\right) \tag{5}$$

$$\mathbf{Z}_{\text{L-1}} = \left[Z^0_{L-1}; Z^{A_1}_{L-1}, Z^{A_2}_{L-1}, \cdots, Z^{A_K}_{L-1}\right] \tag{6}$$

### *3.3. Multi-Stream Transformer Block*

As different deepfake generation techniques target different proportions and regions of the facial area to be manipulated, ranging from small regions such as color mismatch in lips to larger areas that extend throughout the image like face boundaries in face-swapping approaches, it is critically important to identify those regions and extract discriminative features in a scalable manner. Therefore, the capability to have a flexible field of view would provide better information compared to a fixed field of view. While the majority of existing literature focuses on only on a fixed field of view, i.e., patch size, this study proposes a multi-level patch extraction and fusion mechanism that can leverage deepfake traces ranging from more significant facial markers, e.g., eyes, nose, and lips, to more subtle details such as the eye's Iris. As depicted in Figure 1, the proposed multiscale deepfake detection framework consists of two branches, including a low-level patch branch and a high-level patch branch, each composed of three main components: patch/positional embedding, a transformer block, and an attention-based patch selection module followed by a multiscale deepfake classifier. The key difference between low-level branch and high-level branch is the size of image patches and how the sequence patch embeddings are constructed from those patch embeddings and positional embeddings. While the low-level transformer block learns from larger numbers of extracted patches in smaller sizes, the high-level transformer block learns more global features from larger image patches. These characteristics of low-level and high-level transformer blocks enable them to efficiently extract local and global features, respectively. For a given image, each branch's extracted patch/positional embeddings will be fed into the corresponding transformer block, i.e., the low-patch transformer block or the high-patch transformer block. As can be seen in Figure 3, each transformer block comprises three residual transformer blocks with three consecutive vision transformer encoders. The intuition behind using the residual connection between adjacent transformer blocks is to extract additional texture features.

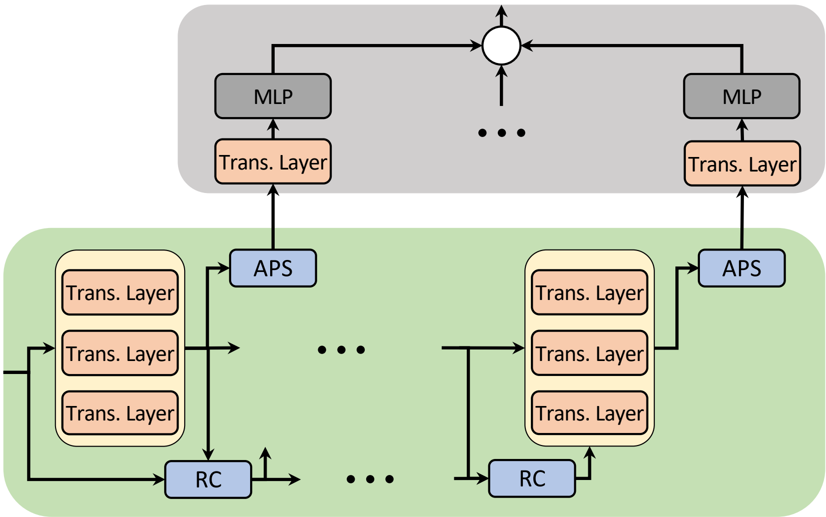


**Figure 3.** Each transformer block comprises several residual transformer blocks with three consecutive re-attention transformer encoders. Top key patches are selected using APS before being fed into the multi-scale deepfake detector (RC: Residual connection, and APS: Attention-based patch selection).

It has been shown that the feature maps tend to be identical in the top layers of deep vision transformer models. This means that the self-attention mechanism fails to learn effective concepts for representation learning. In other words, traditional multi-head self-attention layers suffer from an attention collapse problem, which prevents the vision transformer model from scaling up and hurts the model performance [54]. Unlike other vision transformer frameworks where each transformer encoder is composed of a multi-head self-attention layer and feed-forward multilayer perceptron, the presented transformer encoder in this study is composed of a re-attention mechanism along with a feed-forward multilayer perceptron. The re-attention mechanism re-generates the attention maps through establishing cross-head communications in an attempt to increase the diversity of attention maps at different layers. The intuition behind the re-attention mechanism is that, while similarity between attention maps across different transformer blocks is high, their similarity from different heads of the same transformer block is small. The general architecture of the traditional transformer layer with the self-attention mechanism and the transformer layer with the re-attention mechanism are demonstrated in Figure 4. Mathematical representation of the traditional multi-head self attention layer and re-attention mechanism can be written as (7) and (8), respectively [54]. Both methods generate a trainable associate memory with a query $Q$ and a pair of key $K$-value $V$ pairs to an output via linearly transforming the input.

$$\text{Attention }(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \tag{7}$$

$$\text{Re} - \text{Attention}(Q, K, V) = \text{Norm}\left(\theta^{\top}\left(\text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)\right)\right)V, \tag{8}$$

here, $\sqrt{d}$ and $\theta$ are a scaling factor based on the depth of the network and a learnable transformation matrix, respectively, whereas Norm is a normalization function.

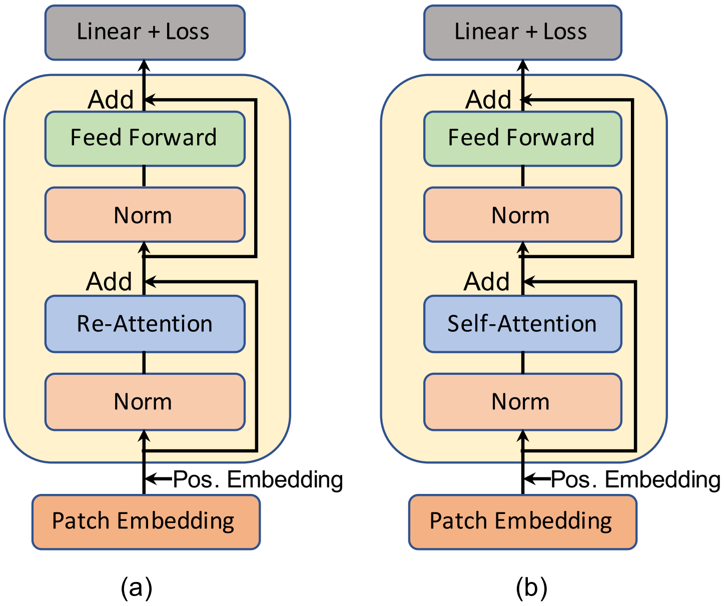


**Figure 4.** Transformer layer with re-attention mechanism vs. self-attention approach [54]. While traditional transformer layers with self-attention mechanisms suffer from the attention collapse problem, the transformer layer with the re-attention mechanism has better scalability. (**a**) Transformer layer with re-attention mechanism; (**b**) Transformer layer with self-attention mechanism.

### 3.4. Multi-Scale Deepfake Detector

Different deepfake generation techniques target different proportions and regions of the facial area in the forgery process, ranging from small regions such as color mismatch in lips to larger areas that extend throughout the image, like face boundaries in face-swapping approaches. Therefore, key patches at different scales may contain plethora of discriminative information that needs to be participated in the decision making process. To do so, the multi-scale deepfake detector takes the following steps, as depicted in Figure 3: (1) making initial prediction based on sequence output of each residual transformer blocks at low/high-level transformer block, and (2) averaging over all prediction from different scales to achieve final decision.

## 4. Evaluation Settings

A detailed description of general evaluation settings, e.g., deepfake datasets, preprocessing steps, implementation specifics, and evaluation metrics, are provided in this section.

### *4.1. Datasets*

Early deepfake forensics benchmarks have significantly enhanced the community's awareness of deepfake threats and encouraged the development of different detection tools. Nonetheless, they suffer from the following drawbacks (1) limited scenes in original videos, (2) low-quality synthesized faces, (3) visible splicing boundaries, (4) color mismatch, (5) visible parts of the original face, and (6) inconsistent synthesized face orientations [55,56]. On the other hand, as adversarial entities are constantly devising new techniques to create more decisive deepfakes, forensics tools must be comprehensive and applicable to challenging real-world applications. Therefore, every dataset selected for evaluation should hold specific characteristics, including diversified real-world scenes, minimal visual artifacts, super-realism and stealth, and covering a wide range of manipulation techniques.

To satisfy this requirement, DFDT's performance is empirically examined against a wide range of high-quality yet challenging benchmarks, including FaceForensics++ [49], Celeb-DF (V2) [55] and WildDeepfake [56]. While FaceForensics++ consists in four different types of facial forgery types, Celeb-DF (V2) and WildDeepfake are the most challenging deepfake datasets in real-world scenarios. For each dataset, 80% of the video samples are held out for training purposes, whereas the the rest are equally divided into test and validation sets. Furthermore, the official test set of four other well-known benchmarks, i.e., DeeperForensics [57], Facebook's DeepFake Detection Challenge (DFDC) dataset [58], and FaceShifter [33] are utilized to evaluate the cross-dataset generalization capability of the DFDT. The purpose of this measurement is to demonstrate how well the model would perform on unseen deepfake samples. A brief description of the utilized datasets are presented as follows and associated statistical specifics are listed in Table 1.

**Table 1.** Statistical specifics of the three benchmarks utilized in this study. Holding diversified real-world scenes, minimal visual artifacts, super-realistic & stealthy [55,56], and covering a wide range of manipulation techniques [49] are key characteristics of the selected benchmarks.

| Dataset | | Statistics | | | | | Source |
|---|---|---|---|---|---|---|---|
| | | Videos | Frames | Train | Test | Val. | |
| FaceForensics++ [49] | Real | 1000 | 509.9 K | 800 | 100 | 100 | YouTube |
| | Deepfake | 4000 | 1830.1 K | 3200 | 400 | 400 | DF |
| Celeb-DF (V2) [55] | Real | 590 (+300) [1] | 225.4 K | 632 | 62 | 196 | YouTube |
| | Deepfake | 5639 | 2116.8 K | 4736 | 536 | 340 | DF |
| WildDeepfake [56] | Real | 3805 | 680 K | 3044 | 380 | 381 | Internet |
| | Deepfake | 3509 | 500 K | 2807 | 350 | 351 | Internet |

[1] Celeb-real plus 300 additional videos downloaded from YouTube [55].

**FaceForensics++.** FaceForensics++ is one of the well-known deepfake detection datasets consisting of four different types of manipulation techniques, including Deepfakes [27], FaceSwap [59], Face2Face [3], and NeuralTextures [60]. It has 1000 real videos from YouTube and corresponding deepfake videos generated using the aforementioned techniques.

**Celeb-DF (V2).** The Celeb-DF (V2) is composed of large-scale deepfake videos generated using an improved synthesis process that swaps faces of individuals in target and source videos. The Celeb-DF (V2) offers high visual quality scores and consists of 5639 deepfakes corresponding to over 2 million frames.

**WildDeepfake.** Unlike FaceForensics++ and Celeb-DF (V2), the WildDeepfake dataset comprises real-world videos for both original and fake videos gathered from the Internet. They are not generated using AI-enabled methods, making them more challenging and closer to real-world scenarios. Furthermore, more diversified scenes, more individuals in each scene, and several facial expressions are among other characteristics of this benchmark.

### *4.2. Implementation Specifics*

A detailed description of the characteristics and technical specifics on the implementation of the proposed method is provided below.

**Implementation.** All models are implemented using the *PyTorch* machine learning library and trained using *Adam* optimizer with a learning rate of $10^{-4}$ with ten times decay every 40 steps. The whole network is trained for 100 epochs.

**Experimental Setup.** Two Lambda Quad deep learning workstation machines were used to conduct all experiments. Each of these machines is installed with Ubuntu 18.04 OS, along with Intel Xeon E5-1650 v4 CPUs, 64 GB DDR4 RAM, 2TB SSD, 4TB HDD, and 4 NVIDIA Titan-V GPUs.

### *4.3. Evaluation Metrics*

The performance of the proposed deepfake detection method is evaluated on both frame-level and video-level analysis. The results are reported using accuracy score (ACC) and/or area under the receiver operating characteristic curve (AUC). These two evaluation metrics commonly have been used in existing deepfake detection tasks [17,61,62]. Therefore, to provide a better understanding and insight into the performance of the presented model, the same metrics are employed in this study.

## 5. Results & Discussion

A comprehensive set of experiments is conducted to evaluate the proposed transformer-based deepfake detection method's performance from various perspectives. The aforementioned evaluation metrics, i.e., detection accuracy, AUC, and recall scores, are employed to measure the performance of the DFDT. Most of the existing deepfake detection methods conduct only frame-level analysis. However, it is critically important to conduct a video-level examination since most deepfake data dissemination on the digital media are forged videos. Therefore, all experiments in this study are performed on two levels, covering both frame-level and video-level. A comprehensive set of experiments are designed to examine the performance of the proposed approach from several aspects. First, the intra-dataset performance of the DFDT is evaluated against three well-known benchmark datasets, including Celeb-DF (V2), WildDeepfake, and FaceForensics++ [49,55,56].

Considering the critical role of the generalization property in deepfake detection task, another set of experiments are designed to examine the cross-dataset generalization capability of the DFDT. Third, the model's performance is compared with that of existing state-of-the-art deepfake detection methods. Finally, the impact of the different components of the DFDT on its function is investigated through various ablation studies. Each of these experiments is discussed in more detail in the following.

### *5.1. Intra-Dataset Evaluation*

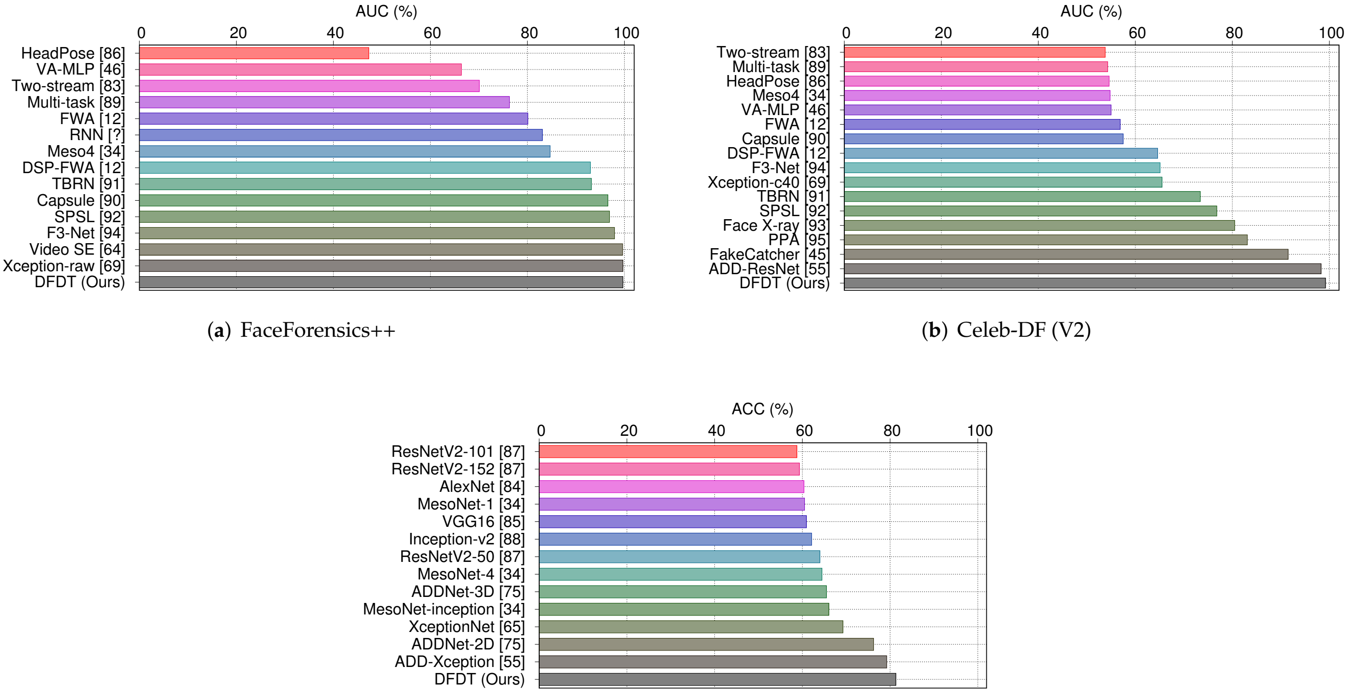
The main goal of this section is to investigate the learning capability of the model and see how well it performs against datasets with different visual qualities and challenging real-world deepfake datasets. Therefore, the model is trained and tested on a range of deepfake detection datasets, spanning different levels of visual quality scores, namely Celeb-DF (V2), WildDeepfake, and FaceForensics++ [49,55,56]. The findings of this experiment demonstrated that DFDT performs significantly well on every challenging dataset on all measured scores. Quantitative frame-level detection results are summarized in Table 2. Particularly, in the frame-level setting the DFDT model has achieved 99.41%, 99.31%, and 81.35% on an accuracy score corresponding to FaceForensics++ (raw) [49], Celeb-DF (V2) [55], and WildDeepfake [56], respectively. A similar trend is apparent in video-level analysis, providing additional strong evidence regarding the outstanding performance of the proposed transformer-based deepfake detection approach in intra-dataset settings. Furthermore, a quantitative comparison of the existing deepfake detection methods with DFDT on every dataset is presented in Table 3. Note that the same evaluation metrics as the literature are utilized for each dataset. It can be observed from Figure 5 that the presented deepfake detection approach in this work outperforms existing methods on all three benchmarks. This figure also reveals another critical point: although most deepfake detection approaches perform well on relatively more straightforward datasets, i.e., FaceForensics++, their performance is still far from perfect on more challenging and real-world datasets, i.e., WildDeepfake.

**Table 2.** Quantitative detection results on different deepfake forensics benchmarks.

| **Dataset ↓** | **Metrics ⟶** | **ACC (%)** | **AUC (%)** |
|---|---|---|---|
| FaceForensics++ | (LQ) | 93.67 | 94.48 |
| | (HQ) | 98.18 | 99.26 |
| | (Raw) | 99.41 | 99.94 |
| Celeb-DF (V2) | | 99.31 | 99.26 |
| WildDeepfake | | 81.35 | 80.74 |

**Table 3.** A quantitative comparison of DFDT's performance on every dataset with existing deepfake detection approaches in frame-level analysis. Reported results are obtained from associated articles. The same evaluation metric as the literature is used for each dataset to provide a fair comparison and better insight into the model's performance.

| **Methods ↓** | **FaceForensics++ AUC (%)** | **Methods ↓** | **Celeb-DF (V2) AUC (%)** | **Methods ↓** | **WideDeepfake (AR %)** |
|---|---|---|---|---|---|
| Two-stream [63] | 70.1 | Two-stream [63] | 53.8 | AlexNet [64] | 60.37 |
| Meso4 [35] | 84.7 | Meso4 [35] | 54.8 | VGG16 [65] | 60.92 |
| HeadPose [10] | 47.3 | HeadPose [10] | 54.6 | ResNetV2-50 [66] | 63.99 |
| FWA [12] | 80.1 | FWA [12] | 56.9 | ResNetV2-101 [66] | 58.73 |
| VA-MLP [34] | 66.4 | VA-MLP [34] | 55.0 | ResNetV2-152 [66] | 59.33 |
| Xception-raw [49] | 99.7 | Xception-c40 [49] | 65.5 | Inception-v2 [67] | 62.12 |
| Multi-task [68] | 76.3 | Multi-task [68] | 54.3 | MesoNet-1 [35] | 60.51 |
| Capsule [69] | 96.6 | Capsule [69] | 57.5 | MesoNet-4 [35] | 64.47 |
| DSP-FWA [12] | 93 | DSP-FWA [12] | 64.6 | MesoNet-inception [35] | 66.03 |
| TBRN [70] | 93.2 | TBRN [70] | 73.4 | XceptionNet [47] | 69.25 |
| SPSL [71] | 96.94 | Face X-ray [72] | 80.5 | ADDNet-2D [56] | 76.25 |
| F3-Net [73] | 97.97 | SPSL [71] | 76.8 | ADDNet-3D [56] | 65.5 |
| Video SE [46] | 99.64 | F3-Net [73] | 65.1 | ADD-Xception [18] | 79.23 |
| RNN [74] | 83.10 | PPA [75] | 83.1 | | |
| | | DefakeHop [6] | 90.5 | | |
| | | FakeCatcher [15] | 91.5 | | |
| | | ATS-DE [7] | 97.8 | | |
| | | ADD-ResNet [18] | 98.3 | | |
| DFDT (Ours) | 99.7 | DFDT (Ours) | 99.2 | DFDT (Ours) | 81.3 |



(**a**) FaceForensics++

(**b**) Celeb-DF (V2)

(**c**) WildDeepfake

**Figure 5.** A quantitative compression of the DFDT approach with existing deepfake detection methods on frame-level analysis. In line with the literature, AUC (%) is used to compare different approaches on FaceForensics++ and Celeb-DF (V2) in (**a**,**b**), respectively, whereas ACC(%) is utilized on WildDeepfake in (**c**). FaceForensics++ is a relatively easier benchmark than Celeb-DF (V2) and WildDeepfake, which have high visual quality scores and are closer to challenging real-world scenes.

### *5.2. Cross-Dataset Generalization*

While deepfake generation methods are constantly evolving and span across classical and AI-driven approaches, it is critically important for any deepfake detection method to have a high generalization capability to recognize unseen samples effectively. To measure this property, in line with existing literature, the DFDT model is trained on FaceForensics++ and then examined on other datasets, including Celeb-DF (V2), DeepFake Detection Challenge, Faceshifter, and DeeperForensics. Table 4 presents the attained AUC scores for video-level analysis. Comparison of the obtained results from this study and the existing state-of-the-art methods demonstrates the excellent cross-dataset generalization capability of the DFDT method. It can be observed from Figure 6 that the DFDT approach achieves remarkable cross-data generalization. It surpasses other systems on Celeb-DF, Faceshifter, and DFDC datasets with relatively large margins and achieves competitive scores on DeeperForensics with the state-of-the-art method, i.e., LipForensics [17].

### *5.3. Cross-Manipulation Generalization*

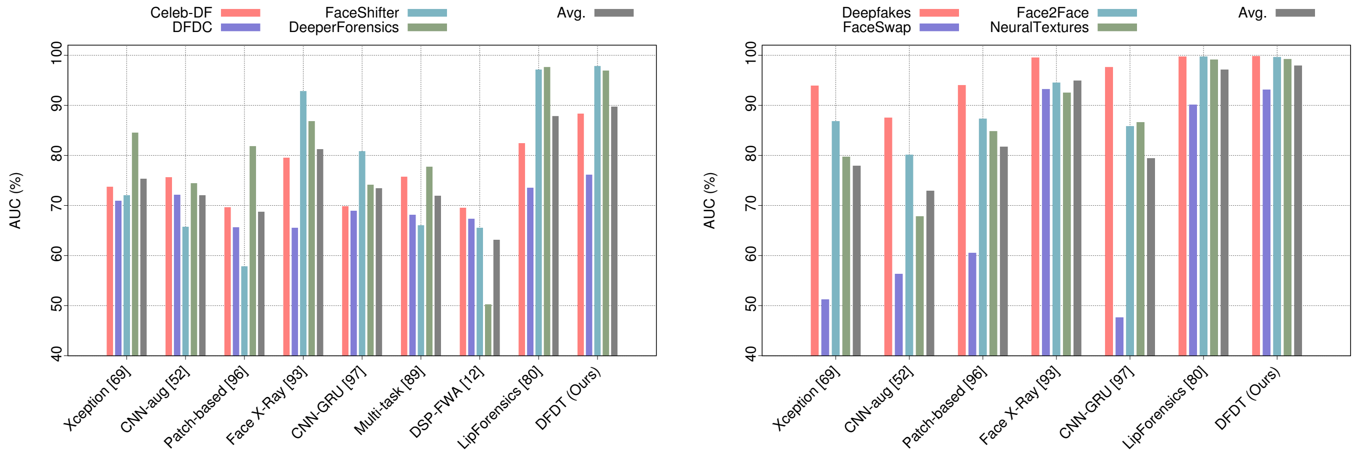
Another set of experiments is designed to understand the generalization capability of the DFDT method to other fake videos created with different manipulation techniques on the same source video while maintaining the pose and illumination variables intact. To do so, DFDT is separately trained on three out of four deepfake generation methods on FaceForensics++ and tested on the remaining one. With the same rationale as in [17,68,70,76], the experiments are performed using the high-quality subset of the FaceForensics++ dataset. It is more likely to be closer to real-world deepfake videos, i.e., where videos are processed with nearly lossless compression. The obtained results from this experiment, as shown in Table 5, show that DFDT's generalization property is well-extended to previously unseen forgery types. Specifically, it achieves higher or competitive cross-manipulation generalization ability compared to existing approaches. As it can be observed in Figure 6, on average, DFDT provides better or competitive scalability to unseen forgery types compared to existing deepfake detection methods.

**Table 4.** Quantitative video-level cross-dataset generalization results (AUC (%)) on Celeb-DF (V2), DeepFake Detection Challenge (DFDC), FaceShifter, and DeeperForensics when trained on FaceForensics++. Reported results in rows 1–9 are from [17].

| **Methods ↓** | **Celeb-DF** | **DFDC** | **FaceShifter** | **DeeperForensics** | **Avg** |
|---|---|---|---|---|---|
| Xception [49] | 73.7 | 70.9 | 72 | 84.5 | 75.3 |
| CNN-aug [39] | 75.6 | 72.1 | 65.7 | 74.4 | 72 |
| Patch-based [76] | 69.6 | 65.6 | 57.8 | 81.8 | 68.7 |
| Face X-Ray [72] | 79.5 | 65.5 | 92.8 | 86.8 | 81.2 |
| CNN-GRU [77] | 69.8 | 68.9 | 80.8 | 74.1 | 73.4 |
| Multi-task [68] | 75.7 | 68.1 | 66 | 77.7 | 71.9 |
| DSP-FWA [12] | 69.5 | 67.3 | 65.5 | 50.2 | 63.1 |
| Two-branch [70] | 76.7 | - | - | - | - |
| LipForensics [17] | 82.4 | 73.5 | 97.1 | 97.6 | 87.8 |
| DFDT (Ours) | 88.3 | 76.1 | 97.8 | 96.9 | 89.7 |

**Table 5.** Video-level cross-manipulation generalization results (AUC (%)) on each subset of FaceForensics++ dataset, including Deepfakes, FaceSwap, Face2Face, and NeuralTextures.

| **Methods ↓** | **Train DFDT on Other Three Subsets** | | | | **Avg.** |
|---|---|---|---|---|---|
| | **Deepfakes** | **FaceSwap** | **Face2Face** | **NeuralTextures** | |
| Xception [49] | 93.9 | 51.2 | 86.8 | 79.7 | 77.9 |
| CNN-aug [39] | 87.5 | 56.3 | 80.1 | 67.8 | 72.9 |
| Patch-based [76] | 94 | 60.5 | 87.3 | 84.8 | 81.7 |
| Face X-ray [72] | 99.5 | 93.2 | 94.5 | 92.5 | 94.9 |
| CNN-GRU [77] | 97.6 | 47.6 | 85.8 | 86.6 | 79.4 |
| LipForensics [17] | 99.7 | 90.1 | 99.7 | 99.1 | 97.1 |
| DFDT (Ours) | 99.8 | 93.1 | 99.6 | 99.2 | 97.9 |



(**a**) Cross-dataset generalization.

(**b**) Cross-manipulation generalization.

**Figure 6.** Examining the generalization capability of the presented end-to-end transformer-based deepfake detection method. (**a**) shows the cross-dataset generalization capability of the DFDT, whereas, the cross-manipulation generalization property on video-level is demonstrated in (**b**).

### *5.4. Ablation Study*

The main goal of this experiment is to examine the impact of different attention mechanisms on the performance of the proposed deepfake detection framework. Two different attention mechanisms, i.e., self-attention and re-attention mechanisms are investigated in this study. The re-attention mechanism re-generates the attention maps through establishing cross-head communications in an attempt to increase the diversity of attention maps at different layers.

Different experiments are conducted with and without such a mechanism to explore the re-attention transformer layer's impact on DFDT's performance. The comparison results on the AUC score are demonstrated in Figure 7. It can be observed from this figure that without the re-attention mechanism the performance of the DFDT decreases by 1.7% and 0.9% and 1.4% in FaceForensics++, Celeb-DF (V2), and WildDeepfake, respectively.

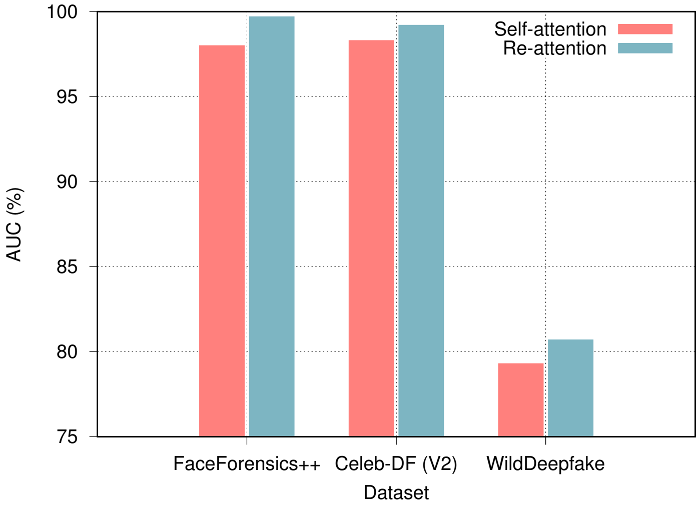


**Figure 7.** Investigating the impact of transformer layer's attention mechanism on the overall performance of the DFDT. The re-attention mechanism has improved the performance of the deepfake detection model compared to self-attention mechanism in transformer layer.

## 6. Conclusions

This work introduces DFDT, an end-to-end deepfake detection framework using vision transformers. Unlike mainstream deepfake detection methods, which exploit CNNs as their backbone, DFDT leverages the unique characteristics of vision transformer networks to model local image features and global relationships of pixels simultaneously. DFDT's multi-stream design enables it to capture different scales of alterations effectively. Obtained experimental results on several benchmarks demonstrate that DFDT achieves state-of-the-art performances, achieving 99.41%, 99.31%, and 81.35% on FaceForensics++, Celeb-DF (V2), and WildDeepfake, respectively. Furthermore, DFDT's excellent cross-dataset & cross-manipulation generalization provides additional strong evidence of its effectiveness.

**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:** All datasets used in this study, i.e., FaceForensics++, Celeb-DF (V2), and WildDeepFake were collected from public Internet and YouTube videos so no consents are obtained.

**Data Availability Statement:** FaceForensics++, Celeb-DF (V2), and WildDeepFake datasets are available upon request at [https://github.com/ondyari/FaceForensics, accessed on 19 August 2020], [https://github.com/yuezunli/celeb-deepfakeforensics, accessed on 5 August 2020] and [https://github.com/deepfakeinthewild/deepfake-in-the-wild, accessed on 12 June 2020], respectively.

**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*. Available online: https://arxiv.org/abs/1406.2661 (accessed on 7 February 2022)
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, 26 June 26–1July 2016; pp. 2387–2395.
4. 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]
5. 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*, 2056305120903408. [CrossRef]
6. Chen, H.S.; Rouhsedaghat, M.; Ghani, H.; Hu, S.; You, S.; Kuo, C.C.J. DefakeHop: A Light-Weight High-Performance Deepfake Detector. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Virtual, 5–9 July 2021; pp. 1–6.
7. Tran, V.N.; Lee, S.H.; Le, H.S.; Kwon, K.R. High Performance deepfake video detection on CNN-based with attention target-specific regions and manual distillation extraction. *Appl. Sci.* **2021**, *11*, 7678. [CrossRef]
8. Shelke, N.A.; Kasana, S.S. A comprehensive survey on passive techniques for digital video forgery detection. *Multimed. Tools Appl.* **2020**, *80*, 6247–6310. [CrossRef]
9. Mirsky, Y.; Lee, W. The creation and detection of deepfakes: A survey. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–41. [CrossRef]
10. 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.
11. 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.
12. 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, 16–20 June 2019; pp. 46–52.
13. 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, Virtual, 19–23 October 2020; pp. 325–334.
14. Jain, A.; Majumdar, P.; Singh, R.; Vatsa, M. Detecting GANs and retouching based digital alterations via DAD-HCNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 672–673.
15. Ciftci, U.A.; Demir, I.; Yin, L. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef]
16. 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.
17. Haliassos, A.; Vougioukas, K.; Petridis, S.; Pantic, M. Lips Don't Lie: A Generalisable and Robust Approach To Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 Jun 2021; pp. 5039–5049.
18. Khormali, A.; Yuan, J.S. ADD: Attention-Based DeepFake Detection Approach. *Big Data Cogn. Comput.* **2021**, *5*, 49. [CrossRef]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
20. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
21. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef]
22. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations, Virtual, 26 April–1 May 2020.
23. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Online, 23–28 August 2020; pp. 213–229.

24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual, 26 April–1 May 2020.
25. Heo, Y.J.; Choi, Y.J.; Lee, Y.W.; Kim, B.G. Deepfake Detection Scheme Based on Vision Transformer and Distillation. *arXiv* **2021**, arXiv:2104.01353.
26. Wodajo, D.; Atnafu, S. Deepfake Video Detection Using Convolutional Vision Transformer. *arXiv* **2021**, arXiv:2102.11126.
27. Faceswap. Faceswap: Deepfakes Software for All. Available online: https://github.com/deepfakes/faceswap (accessed on 7 February 2022).
28. FakeApp. *FakeApp 2.2.0-Download for PC Free*. Available online: https://www.malavida.com/en/soft/fakeapp/ (accessed on 7 February 2022).
29. 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.
30. 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, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1857–1865.
31. 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.
32. 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]
33. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv* **2019**, arXiv:1912.13457.
34. 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 Village, HI, USA, 7–11 January 2019; pp. 83–92.
35. 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.
36. 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.
37. 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.
38. 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]
39. Wang, S.Y.; Wang, O.; Zhang, R.; Owens, A.; Efros, A.A. CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 8695–8704.
40. 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, Seattle, WA, USA, 2–16 October 2020; pp. 2823–2832.
41. Quan, R.; Wu, Y.; Yu, X.; Yang, Y. Progressive transfer learning for face anti-spoofing. *IEEE Trans. Image Process.* **2021**, *30*, 3946–3955. [CrossRef]
42. Floridi, L.; Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]
43. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision transformers for remote sensing image classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]
44. Yang, F.; Zhai, Q.; Li, X.; Huang, R.; Luo, A.; Cheng, H.; Fan, D.P. Uncertainty-guided transformer reasoning for camouflaged object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4146–4155.
45. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
46. Khan, S.A.; Dai, H. Video Transformer for Deepfake Detection with Incremental Learning. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 1821–1828.
47. 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.
48. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
49. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 1–11.

50. 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, 14–19 June 2020; pp. 5203–5212.
51. 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.
52. Abnar, S.; Zuidema, W. Quantifying attention flow in transformers. *arXiv* **2020**, arXiv:2005.00928.
53. He, J.; Chen, J.N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C.; Yuille, A. TransFG: A Transformer Architecture for Fine-grained Recognition. *arXiv* **2021**, arXiv:2103.07976.
54. Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; Feng, J. Deepvit: Towards deeper vision transformer. *arXiv* **2021**, arXiv:2103.11886.
55. 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.
56. 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, Seattle, WA, USA, 12–16 October 2020; pp. 2382–2390.
57. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2889–2898.
58. Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Ferrer, C.C. The deepfake detection challenge (dfdc) preview dataset. *arXiv* **2019**, arXiv:1910.08854.
59. Faceswap. Faceswap. Available online: https://github.com/MarekKowalski/FaceSwap/ (accessed on 7 February 2022).
60. Thies, J.; Zollhöfer, M.; Nießner, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–12. [CrossRef]
61. Wang, J.; Wu, Z.; Chen, J.; Jiang, Y.G. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. *arXiv* **2021**, arXiv:2104.09770.
62. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2185–2194.
63. 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.
64. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
65. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representation, Lisbon, Portugal, 1–15 September 2015.
66. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2016; pp. 630–645.
67. 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.
68. 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.
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 Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 667–684.
71. Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; Yu, N. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 772–781.
72. 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.
73. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 86–103.
74. 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.
75. 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, 16–17 October 2020.
76. Chai, L.; Bau, D.; Lim, S.N.; Isola, P. What makes fake images detectable understanding properties that generalize. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 103–120.
77. 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.