electronics

MDPI

Article

Improving Detection of DeepFakes through Facial Region Analysis in Images

Fatimah Alanazi [1,2,*], **Gary Ushaw** [1] **and Graham Morgan** [1]

[1] School of Computing, Newcastle University, Newcastle upon Tyne NE1 7RU, UK ; gary.ushaw@newcastle.ac.uk (G.U.); graham.morgan@newcastle.ac.uk (G.M.)
[2] College of Computer Science and Engineering, University of Hafr Al Batin, Hafar Al Batin 39524, Saudi Arabia
[*] Correspondence: f.m.z.alanazi2@newcastle.ac.uk

Abstract: In the evolving landscape of digital media, the discipline of media forensics, which encompasses the critical examination and authentication of digital images, videos, and audio recordings, has emerged as an area of paramount importance. This heightened significance is predominantly attributed to the burgeoning concerns surrounding the proliferation of DeepFakes, which are highly realistic and manipulated media content, often created using advanced artificial intelligence techniques. Such developments necessitate a profound understanding and advancement in media forensics to ensure the integrity of digital media in various domains. Current research endeavours are primarily directed towards addressing a common challenge observed in DeepFake datasets, which pertains to the issue of overfitting. Many suggested remedies centre around the application of data augmentation methods, with a frequently adopted strategy being the incorporation of random erasure or cutout. This method entails the random removal of sections from an image to introduce diversity and mitigate overfitting. Generating disparities between the altered and unaltered images serves to inhibit the model from excessively adapting itself to individual samples, thus leading to more favourable results. Nonetheless, the stochastic nature of this approach may inadvertently obscure facial regions that harbour vital information necessary for DeepFake detection. Due to the lack of guidelines on specific regions for cutout, most studies use a randomised approach. However, in recent research, face landmarks have been integrated to designate specific facial areas for removal, even though the selection remains somewhat random. Therefore, there is a need to acquire a more comprehensive insight into facial features and identify which regions hold more crucial data for the identification of DeepFakes. In this study, the investigation delves into the data conveyed by various facial components through the excision of distinct facial regions during the training of the model. The goal is to offer valuable insights to enhance forthcoming face removal techniques within DeepFake datasets, fostering a deeper comprehension among researchers and advancing the realm of DeepFake detection. Our study presents a novel method that uses face cutout techniques to improve understanding of key facial features crucial in DeepFake detection. Moreover, the method combats overfitting in DeepFake datasets by generating diverse images with these techniques, thereby enhancing model robustness. The developed methodology is validated against publicly available datasets like FF++ and Celeb-DFv2. Both face cutout groups surpassed the Baseline, indicating cutouts improve DeepFake detection. Face Cutout Group 2 excelled, with 91% accuracy on Celeb-DF and 86% on the compound dataset, suggesting external facial features' significance in detection. The study found that eyes are most impactful and the nose is least in model performance. Future research could explore the augmentation policy's effect on video-based DeepFake detection.

Keywords: DeepFake detection; face augmentation; face cutout facial recognition; feature fusion; image analysis

check for updates

Citation: Alanazi, F.; Ushaw, G.; Morgan, G. Improving Detection of DeepFakes through Facial Region Analysis in Images. *Electronics* **2024**, *13*, 126. https://doi.org/10.3390/electronics13010126

Academic Editors: Yue Wu and Honarvar Shakibaei Asli

Received: 20 November 2023
Revised: 20 December 2023
Accepted: 25 December 2023
Published: 28 December 2023

1. Introduction

The recent increase in affordable intelligent devices, including digital cameras, laptops, tablets, and cell phones, has led to a significant rise in digital multimedia content production, notably audio, images, videos, and text [1]. These devices, equipped with advanced operating systems, facilitate the use of applications that can alter multimedia content, contributing to the emergence of a "post-truth era", where factual information is often replaced with alternative narratives [2].

The surge in digital multimedia content on social media has led to a variety of fake content, highlighting the need for sophisticated detection methods. Deep learning algorithms have become essential in identifying these falsities across various domains. For example, in transportation, the DMF Network [3] uses disparity maps to improve the detection of small objects at long distances. Additionally, a study [4] on Space–Air–Ground Integration Networks employs federated learning for efficient communication and semi-supervised learning to enhance anomaly detection. These examples highlight the crucial role of deep learning in addressing challenges related to detecting information authenticity and ensuring network security.

Highlighting the significance of recent advancements in DeepFake detection, the study [5] underscores the effectiveness of graph–transformer techniques for handling complex datasets with nuanced features. This is particularly pertinent to DeepFake detection, where analysing subtle visual cues with limited data is essential. Furthermore, the challenges in hyperspectral image analysis, such as handling high-dimensional data and limited labeled samples, parallel the complexities in identifying sophisticated DeepFakes.

Building on this, the research in [6] underscores the importance of integrating multiple features for accurately recognising facial expressions. This concept is directly applicable to DeepFake detection, wherethe amalgamation of various facial features and expressions can enhance the accuracy of identifying manipulated videos.The application of Convolutional Neural Networks (CNNs) in [7] demonstrates their versatility and effectiveness across different contexts, including industrial defect recognition. This versatility aligns with their crucial role in DeepFake detection, distinguishing authentic from altered content. Additionally, ref. [8] extends the scope of machine learning applications to pattern recognition and sensor data analysis. These methodologies can inform strategies in DeepFake detection, particularly regarding data processing and pattern identification.

DeepFakes, as a category of altered multimedia, have garnered significant interest recently. These multimedia manipulations, enabled by deep learning, allow for face substitutions in videos, transfer of facial expressions, and alteration of physical and facial characteristics [9]. The field of face manipulation has witnessed a surge in interest in DeepFake techniques, leading to substantial progress in creating manipulated images [10,11].

Although DeepFake videos can be used for entertaining or harmless purposes, they also carry the potential for malevolent motives. They can be manipulated to target individuals and spread false information, and possess the power to manipulate public sentiment and impact election results by spreading deceptive content about political figures [12].

A primary challenge in DeepFake detection is the problem of overfitting. Overfitting arises when a model becomes overly attuned to the training data, leading to poor generalisation to previously unseen data, resulting in subpar performance when encountering real-world DeepFakes [13]. To mitigate the issue of overfitting, data augmentation techniques have been employed. Data augmentation involves the creation of synthetic data from existing data, acting as a means to prevent the model from becoming overly familiar with the training data and, thus, improving its performance with novel data.

A prevalent data augmentation technique in DeepFake detection is random erasure or cutout, which entails randomly removing segments from images. This strategy helps reduce the model's excessive reliance on particular image features, as outlined in Khan et al. [14]. However, the arbitrary nature of this technique may unintentionally conceal essential facial regions crucial for DeepFake detection, potentially diminishing the model's overall effectiveness.

The proposed method, named 'Face-Cutout', offers valuable guidelines for DeepFake detection research. This technique dynamically selects and occludes specific regions of the face, guided by facial landmark information. It enhances the diversity of training data while avoiding the obscuration of critical facial areas. By providing specific guidelines for facial region removal, Face-Cutout aids deep learning models in focusing on and learning from the most informative facial features in an image. This approach addresses the issue of overfitting in DeepFake datasets and improves the efficacy of deep neural models by pinpointing crucial regions rich in significant information, thereby aiding in the detection of DeepFakes.

A state-of-the-art approach was employed by [15], that is, an interpatch dissimilarity estimator and a multi-stream convolutional neural network to capture distinct DeepFake indicators associated with each feature. By leveraging these indicators, we improve the efficiency and applicability of DeepFake detection. Similarly, in [13], the authors introduced a straightforward data augmentation technique called “Face-Cutout”. Our approach involves dynamically excising sections of an image based on facial landmark information. This aids the model in focusing exclusively on the pertinent portions of the input. Likewise, the authors in [16] conducted a comprehensive examination of DeepFake creations, focusing on facial regions and the efficacy of fake detection. Furthermore, in [17], the authors introduced an attention-driven multi-task strategy aimed at enhancing feature maps for classification and localisation objectives. In this network, the encoder and attention-driven decoder work collaboratively to produce localised maps that emphasise areas containing data regarding the nature of the manipulation. These specific characteristics are shared with the classification network, thereby enhancing its efficiency. Rather than relying on encoded spatial attributes, they combined attention-based localised features from the initial layer of the decoder with frequency domain features to establish a distinct and discriminative representation for DeepFake detection.

On the other hand, the authors in [18] proposed a hierarchical and interpretable forensics algorithm that integrates human judgment into the detection process. They utilised a deep learning detection algorithm to curate the data and provide a comprehensible decision to human evaluators, accompanied by a series of forensic analyses pertaining to the decision area. In the domain of detection, we introduce an interpretable DeepFake detection algorithm that relies on attention mechanisms. To overcome the challenge of generalisation, they tackled it by creating an ensemble of detection networks, including both conventional and attention-based models, with data augmentation. However, it is important to note that many of the approaches described above have primarily been designed to identify particular categories of DeepFake datasets. This specialisation often results in reduced performance when confronted with novel and unfamiliar DeepFakes. Additionally, their limitations in generalisation extend to the varying quality of DeepFakes, which is evident from their subpar performance when confronted with compressed or manipulated inputs [19].

The primary challenge in DeepFake detection lies in the propensity of models to overfit training data, resulting in suboptimal performance when confronted with new, unseen images. Overfitting arises when a model excessively adapts to the specific examples on which it was trained, diminishing its capability to effectively counter diverse DeepFake techniques. A prevalent solution in the literature involves data augmentation strategies, notably the application of random erasure or cutout techniques. However, the arbitrary nature of these methods risks obscuring critical facial regions that are essential for accurate DeepFake detection. In the absence of definitive guidelines for specific facial area removal, most studies have adopted a randomised approach.

To address these concerns, our study implements a “Face-Cutout” technique, which entails the random erasure of image segments to introduce variation in the training dataset and counteract overfitting. This method differentiates between altered and original images, preventing the model from overly conforming to specific training samples, thereby bolstering its generalisation capabilities for new data, leading to improved outcomes.

Moreover, this study examines the importance of different facial elements in DeepFake detection by selectively removing facial regions during model training. This approach aims to provide insights for future DeepFake dataset preparation, identifying key facial areas for effective detection. Our goal is to refine training methods and develop stronger models. This study contributes to a better understanding of DeepFake detection, potentially improving model accuracy and complicating the creation of convincing DeepFakes. The use of the Face-Cutout technique is a promising strategy in DeepFake research, and this paper further explores its implications and benefits in enhancing DeepFake detection.

The organisation of this paper is as follows. Section 2 provides a concise review of the relevant literature and related work. Section 3 then explains the methodology employed in this study, encompassing the augmentations utilised such as face cutouts, the specific models employed, the datasets employed for training the models, and details of the implementation procedure. Section 4 presents the results of our experimentation and, finally, Section 5 presents the conclusions of the study and suggests potential avenues for future research.

2. Related Work

A widely used data augmentation technique is random erasure, involving the cutting out or replacement of random image patches with noise during model training. This method enables the network to learn or infer features that align with adjacent regions in the image, thereby enhancing the model's robustness. However, this random removal of patches can lead to the loss of crucial object descriptors, adversely affecting the training process. Consequently, it is essential to judiciously select data augmentation techniques to prevent the loss of important descriptors and avoid introducing bias into the training data. Moreover, monitoring the model's performance throughout the training is vital to identify any issues and modify the augmentation methods as needed.

The emergence of DeepFake technology presents a profound challenge, chiefly manifested in the escalating intricacy in discerning genuine facial representations from fabricated counterparts within media content [9]. This burgeoning dilemma accentuates a disconcerting reality wherein digital falsification is increasingly melding with truth, thereby posing a significant challenge to both individuals and societal entities. The ability to create realistic counterfeit videos and images erodes trust and creates an environment where the line between reality and fabrication becomes increasingly blurred. The implications of this technology are extensive, impacting not only personal engagements but also the broader societal frameworks and institutions [20]. Consequently, the quandary of DeepFake technology demands prompt attention and robust solutions to alleviate its detrimental effects on the sanctity of digital communication.

Existing methodologies for detecting DeepFakes employ a variety of detection clues to identify such digital falsifications. Among these, Anomalies models are noteworthy, capitalising on the artefacts engendered during the DeepFake creation process to identify counterfeit digital content. Numerous strategies have been harnessed to detect DeepFakes, with a prominent role attributed to convolutional neural networks [21]. Concurrently, Biometric clues represent another avenue, leveraging various human traits for the detection and recognition of fabricated content, particularly focusing on facial features [22]. This approach has witnessed a surge in utilisation recently and is the focal point of examination in this study, providing a nuanced understanding of its efficacy and potential advancements in mitigating the challenges posed by DeepFake technology.

Numerous spoofing attacks have surfaced and been employed with malicious intent, often leading to social discord. The process of individual identification and authentication has been greatly enhanced by the development and implementation of biometric systems.These systems have been widely adopted by various entities, from international and national organisations to individual users, becoming a fundamental aspect of security measures.

Research spearheaded by Jung et al. [23] honed in on the natural blinking of eyes as a pivotal clue for detecting DeepFake videos, employing an algorithm coined as Deep Vision. Given the predictable patterns inherent to eye-blinking—a spontaneous and voluntary action—the technique was applied to a set of eight videos, successfully detecting DeepFake content in seven instances. Nevertheless, the exploration of combined cues remains an under-charted territory despite the encouraging results garnered from single cue models.

Supplementing this, studies such as the one conducted by Menotti et al. [24] pivoted toward the eye region, blending two approaches: the employment of a suitable convolutional network architecture, and the revision of network weightings through back-propagation. Utilising facial, iris, and fingerprint modalities for spoofing detection, they leveraged discrepancies in iris movement—characteristically delayed in counterfeit videos—as a distinctive clue for differentiation between authentic and forged videos. The study underscored the remarkable efficacy of these methodologies in addressing nine identified problem areas.

However, a significant limitation is the reliance on a single indicator, namely, the eyes, for DeepFake detection. This approach becomes less effective when the eyes are obscured for any reason, such as occlusions, making it less adaptable to different scenarios. The success of this method in identifying DeepFakes can vary greatly depending on the specific situation and unique characteristics of the video. Despite its high effectiveness in detecting deep learning-generated fake videos and images, this model depends entirely on the absence of eye-blinking as its main criterion. This can be a vulnerability, as it might not work in videos with frequent eye-blinking, faces altered to have closed eyes during training, or when forgers create realistic blinking.

Furthering the discourse, Ciftci et al. [25] introduced a methodology that leverages biological data to scrutinise videos for forensic augmentations, utilising facial indicators such as heart rate. Employing SVM and CNN models trained on both temporal and spatial facets of facial features, they endeavoured to segregate authentic videos from fabricated ones. Despite enhancing the precision of DeepFake detection, a significant drawback emerges when dimensionality reduction techniques are employed, markedly diminishing the detection accuracy in videos.

Moreover, the proliferation of smartphones equipped with digital camera features and applications enabling digital content editing and sharing, as noted by Jafar et al. [26], alongside the rise of artificial intelligence (AI) via deep learning tools, has enabled the distortion of genuine image and video characteristics for malicious purposes, potentially inciting social unrest. Their research proposed a model employing a convolutional neural network and the DFT–MF technique for detecting fake videos and images. The clues for detection encompassed lip or mouth movements, evaluated through the examination and authentication of datasets derived from the DeepFake Forensics (Celeb-DF) and DeepFake Vid-TIMIT datasets. The methodologies and techniques applied exhibited a high level of accuracy in analysing mouth movements during specific word pronunciations. The scrutiny of fake videos revealed a notably wider and more open mouth in comparison to authentic videos within the datasets. However, a caveat remains that the long-term sustainability of this model may be challenged by the continual advancements in DeepFake creation processes.

Facial occlusion presents a significant challenge in detection tasks, as it diminishes the available information and introduces additional noise into the data. Broadly, facial occlusion can be categorised into two distinct types: landmark occlusion and heavy occlusion. Landmark occlusion is characterised by the obstruction of a few facial landmarks such as eyes or mouths, while the majority of the face remains visible. On the other hand, heavy occlusion entails scenarios where more than half of the face is obscured due to occlusion, image border constraints, or extreme poses, with the challenge exacerbated when the occlusion originates from other faces. Despite the inherent issues that facial occlusion causes in face detection, it has the potential to address overfitting problems common in DeepFake datasets if applied strategically to certain facial areas. This concept forms the basis of our study.

Nonetheless, a myriad of studies has already ventured into leveraging occlusion in various capacities, setting a foundation upon which our investigation is built [27]. The utilisation of random cutout has been acknowledged for enhancing deep models in image classification, object detection, and person re-identification domains [28]. Recently, a surge in the application of cutout techniques has been observed, particularly in digital images embodying faces. A notable study delved into the exploration of face cutout augmentation alongside random cutout augmentations [14]. These augmentations were independently employed to train two distinct models aimed at detecting DeepFakes. The primary objective behind these cutout augmentations is to curb overfitting—a prevalent issue in machine learning where a model overly specialises in training data, thereby faltering in generalising well to novel data. Efforts have been made to ameliorate the generalisation issue within DeepFake detection models, intending to sensitise the proposed model to a diverse array of forgeries through a broad forgery augmentation space [29].

Beyond mitigating overfitting and bolstering generalisation, the face cutout technique in DeepFake detection also serves to construct a DeepFake dataset comprised entirely of masked faces. A study motivated by the COVID-19 pandemic's requisites, aimed to equip DeepFake detection models with the capability to identify manipulated faces even when certain facial regions like the mouth and nose are masked, was performed in [30]. In this endeavour, the authors generated both authentic and counterfeit faces donning masks to construct a test dataset, serving as a basis for evaluating DeepFake detection methodologies. Training the model with both genuine and fake masked faces enhanced its proficiency in identifying DeepFakes with masked regions. However, the study's reliance on a small dataset and testing the model with a dataset having the identical crop as the training dataset hinders the generalizability of the results to other datasets.

An alternative strategy to tackle overfitting in DeepFakes was showcased in a study that employed two cutout operations: sensory group removal and convex-hull removal [13]. Sensory group removal randomly selects one among the three landmark groups—two eyes, the nose, and mouth—and excises the largest polygonal region delineated by the group's points. Conversely, in convex-hull removal, landmark points representing the facial boundary are chosen, and the largest polygon, encapsulated by the points possessing the minimum envelope, is determined. Points may be randomly selected from all boundary points, or as eight or more contiguous points boasting the maximum polygonal area with a minimum envelope.

In conclusion, this study champions a dynamic face augmentation that utilises the Face-Cutout technique to engender a myriad of samples, aiming to mitigate the overfitting dilemma prevalent in DeepFake technology, also this study focuses to elucidate the differences between employing landmark occlusion and heavy occlusion. Nonetheless, the study falls short in furnishing a comprehensive explanation regarding the selection criteria for each type of cutout or identifying which facial features are more information-rich and, thus, should be preserved during the cutout process. This oversight manifests a lack of thorough analysis concerning each Face-Cutout. This gap in analysis potentially hinders a more nuanced understanding and application of the Face-Cutout technique in combating DeepFake propagation. Previous investigations into the Face-Cutout technique have not provided a lucid elucidation regarding the selection criteria for cutout regions, especially concerning facial parts of the images. To bridge this gap, the current research endeavours to unearth which facial features are paramount for effective DeepFake detection.

In summary, our study introduces a dynamic face augmentation method using the Face-Cutout technique to produce diverse samples and tackle the overfitting issue in DeepFake technology. Prior studies have not clearly articulated the rationale behind selecting certain regions for cutout, especially in the context of facial features. Addressing this gap, our research concentrates on identifying the facial features most beneficial for DeepFake detection.

3. Methodology

This study aims to provide crucial insights into enhancing face extraction processes in future DeepFake datasets. In doing so, it seeks to enrich researchers' understanding and advance the field of DeepFake detection, particularly by improving the precision and success rate of detection models. To achieve the study's objectives and identify the most critical facial features for recognising DeepFakes, we have developed a comprehensive methodology. This methodology consists of several key steps, each of which are elaborated upon in the subsequent sections. These steps are visually illustrated in Figure 1, offering an overview of the research methodology employed in this study. The process can be summarised in four stages, starting with face cutout, followed by pre-processing and training, and culminating in testing.

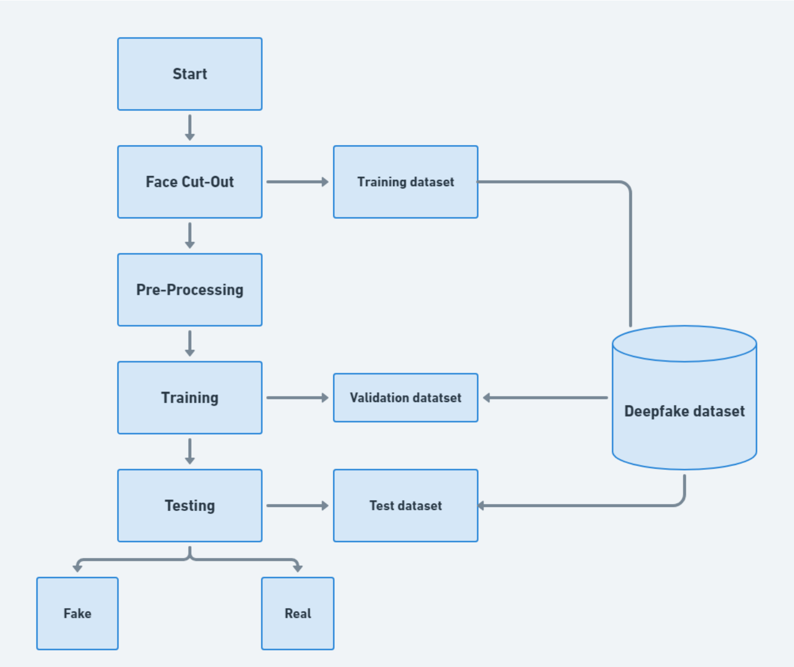


Figure 1. Overview of the research methodology.

3.1. Dataset Selection

In recent years, several DeepFake datasets have been published to facilitate research and development in the field of DeepFake detection. These datasets typically contain a collection of real and fake videos, where the fake videos are generated using various DeepFake techniques such as face swapping and facial re-enactment. Some of the popular DeepFake datasets include the FaceForensics++ (FF++), Celeb-DF, and the DeepFake Detection Challenge (DFDC), and they have been widely used by researchers to develop and evaluate DeepFake detection algorithms and techniques. The models in this study were trained and evaluated with FaceForensics++(FF++) and Celeb-DF, which are presently the most popular datasets from their respective generations. The details are as follows:

- FaceForensics++ datasets [31] is a public benchmark dataset for research in to the detection of face forgery created by researchers from the Technical University of Munich and the University of Erlangen-Nuremberg in Germany and the Federico II University of Naples in Italy. The dataset is characterised by its diverse manipulations and high-quality content, which collectively challenge both human perception and algorithmic detection capabilities. It incorporates manipulations from various methodologies, including DeepFakes, Face2Face, FaceSwap, and Neural Textures, ensuring a robust and comprehensive dataset that spans multiple manipulation techniques. Moreover, the dataset features over 1000 original video sequences, along with their manipulated counterparts, collectively amassing a staggering total of over 5000 videos, all extracted from realistic contexts such as news interviews. FaceForensics++ was created to facilitate research in the area of DeepFake detection and to help develop algorithms that can accurately detect manipulated videos. It has been used in several research papers and challenges, and has become one of the standard benchmarks in its field;

- Celeb-DF dataset [32]—specifically, its second version (Celeb-DF-v2)—has emerged as a pivotal resource in the domain of DeepFake research, boasting a comprehensive collection of 5639 videos, which are meticulously categorised into real and fake. The dataset encompasses videos of 32 distinct celebrities, providing a rich and diverse repository for exploring the intricacies of DeepFake generation and detection. With 590 genuine videos and a staggering 5049 fake videos, Celeb-DF not only presents a broad spectrum of content but also ensures substantial depth, thereby providing researchers with a robust and varied dataset to delve into the complexities of DeepFake technology. It ensures a balanced and realistic framework for developing and testing DeepFake detection algorithms, providing a genuine challenge to researchers and technologists.

Moreover, the dataset is strategically partitioned into training and validation sets, with the training set comprising 4925 videos (468 real and 4457 fake) and the validation set including 714 videos (122 real and 592 fake). This intentional separation facilitates effective model training and evaluation, allowing researchers to navigate the nuanced landscape of DeepFake detection with a structured and methodical approach.

The clear labelling of videos as either real or fake further enables the implementation of supervised learning methodologies, thereby catalysing advancements in the development of robust and efficient DeepFake detection algorithms. Figure 2, in the paper showcases a selection of face samples extracted from the FF++ and DFDC datasets, which were central to this study.

In this study, face cutouts were separately evaluated using the FF++ and Celeb-DF datasets, in addition to training models with samples from both datasets. Generally, when training a machine learning model, the data should be divided into three sets: training, validation, and testing. The training set is utilised to train the model. The validation set, on the other hand, is used to assess the model's performance during training and to aid in selecting hyperparameters. Finally, the testing set is employed to evaluate the model's ultimate performance. For this study, 80% of the data were allocated for training, 10% for validation, and 10% for testing. This distribution is a standard approach, although the specific percentages may vary based on the dataset's size and complexity, among other considerations.

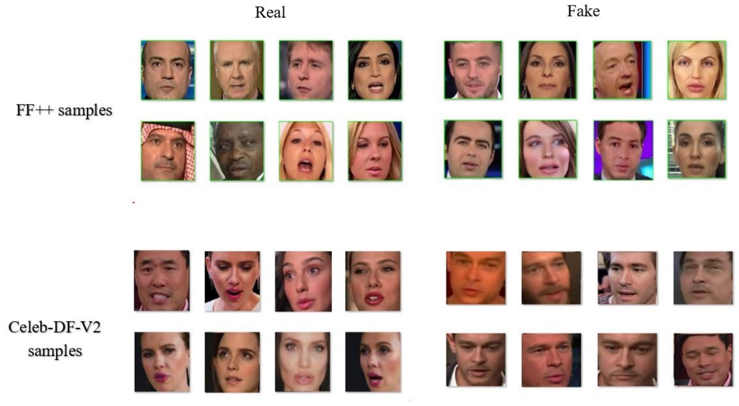


Figure 2. Face samples extracted from FF++ and DFDC datasets.

3.2. Face Cutout

The Face-Cutout technique is acknowledged as an advanced data augmentation strategy used in the training phases of Convolutional Neural Networks (CNNs). It is specifically tailored to enhance the performance of DeepFake detection systems. This technique is instrumental in creating a varied collection of training images characterised by diverse occlusions. It leverages detailed facial landmark data to ensure effectiveness, regardless of the various facial orientations.

These facial landmarks, integral to this approach, denote precise locations of prominent features on the human face, such as the eyes, ears, nose, mouth, jawline, and forehead. A significant instrument in this field, the MediaPipe Face Mesh, is capable of accurately identifying 468 landmark positions on a face. In Figure 3, we can observe the results of the MediaPipe Face Mesh algorithm, which effectively identifies and maps a total of 468 landmark positions on the human face. These positions are scrupulously marked, each representing a point in a sequence from 0 to 468, offering an exhaustive topological insight into facial features that are crucial for augmentation [33].

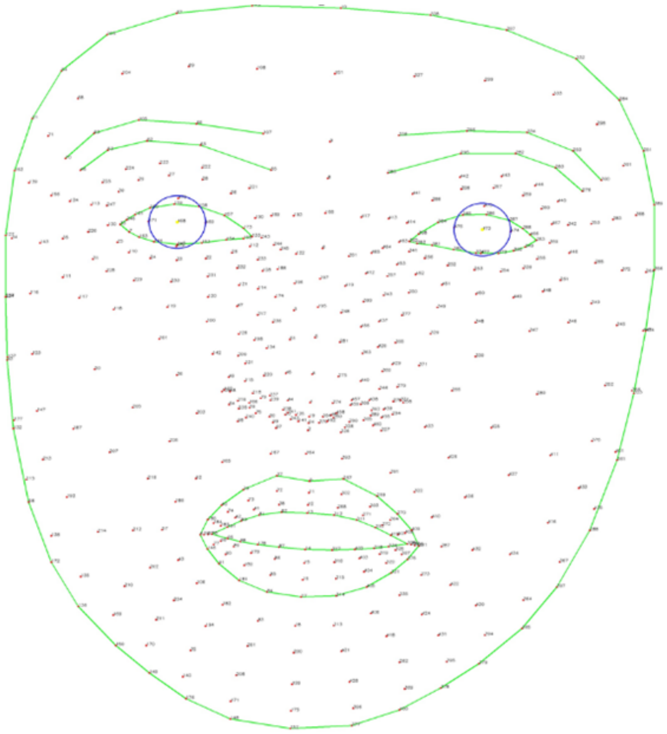


Figure 3. MediaPipe Face Mesh: A 3D Facial Landmark Detector with 468 Landmarks.

MediaPipe Face Mesh distinguishes itself from other facial tracking technologies through its high accuracy, cross-platform compatibility, and user-friendly design, making it an ideal choice for a wide range of applications in both commercial and research settings within the fields of computer vision and interactive technology. Additionally, as an open-source framework, MediaPipe offers developers the flexibility to tailor the Face Mesh model to meet specific requirements. The strong community of developers and users associated with MediaPipe contributes a wealth of shared knowledge, resources, and support, which are invaluable for refining implementations and resolving challenges. Furthermore, the capability of MediaPipe Face Mesh to handle an increased number of landmarks beyond the standard 468 offers the potential for more detailed facial analysis, enhancing accuracy for specific applications, which aligns with the requirements of our study.

Originating from Google's technological suite, MediaPipe Face Mesh stands as a cutting-edge facial landmark detection framework. This state-of-the-art machine learning model harnesses the complexity of deep neural networks to detect and track 468 unique landmark positions on the human face in real-time, delineating essential facial components, including the eyes, eyebrows, nose, mouth, lips, and jawline [33]. The present study incorporates the tactical application of cutout techniques that target specific facial regions, and this experiment was segmented into three phases:

- Phase One: Apply the Face-Cutout technique with each dataset separately. Phase One involved the independent application of the Face-Cutout technique to each dataset (FaceForensics++ and Celeb-DF), resulting in three image sets per dataset: a Baseline group (no cut), Face-Cutout 1 (focusing on core features with cuts on the chin, mouth, jawline, and forehead), and Face-Cutout 2 (emphasising external features with cuts on the left eye, right eye, both eyes, and nose). Subsequent training of deep learning models utilised all groups, followed by performance assessments on unseen datasets;
- Phase Two: Apply the Face-Cutout technique with the combined dataset. Phase Two combined both datasets (FaceForensics++ and Celeb-DF) into a unified set, maintaining the same three-group structure as in Phase One;
- Phase Three: Apply the Face-Cutout technique for Each Facial Feature Separately. Phase Three entailed individualised training with each facial feature obscured separately, followed by performance evaluations to discern the impact of obscured regions on the model's discernment capabilities between authentic and falsified faces. This led to the creation of nine specific training subsets, each corresponding to different facial areas, with a ninth set containing images with no modifications. This strategic segmentation aimed to assess the model's relative performance with and without face cutouts and to identify critical facial areas for efficient model training. Following the group preparations, a comprehensive evaluation phase was conducted to determine each subset's competence in distinguishing synthetic images from genuine ones. Figure 4 showcases examples of the datasets created for this study, which encompass three distinct groups: (1) Baseline images featuring unaltered, original faces; (2) Face-Cutout 1, involving the removal of specific regions like the left eye, right eye, both eyes, and nose; and (3) Face-Cutout 2, which includes the removal of the forehead, chin, mouth, and jawline.

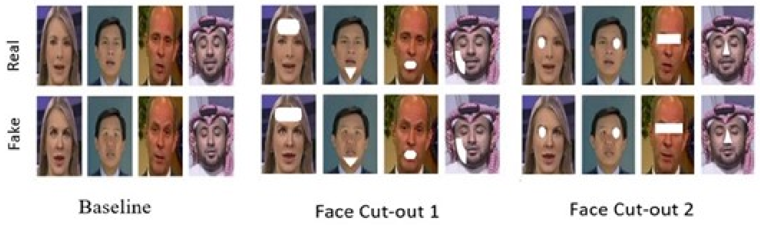


Figure 4. Examples of the datasets generated in this study, comprising three distinct groups: (1) Baseline images representing original, unaltered faces; (2) Face-Cutout 1, which involves cutting out specific regions such as the left eye, right eye, both eyes, and nose; and (3) Face-Cutout 2, which cuts out the forehead, chin, mouth, and jawline.

Overall, the experiment was conducted in three phases, each targeting a unique goal. In Phase One, we applied our approach to two different datasets, Celeb-DF and FF++, separately, both created using various DeepFake generation methods. The objective was to evaluate the effectiveness of our face cutout techniques across these different datasets and check for consistency in the results. Phase Two involved merging the Celeb-DF and FF++ datasets, thereby presenting the model with a broader range of images within the same group and doubling the sample size. This phase aimed to test the model's adaptability and robustness with a larger and more varied image set. Finally, Phase Three shifted focus to the impact of individual facial features on the model's DeepFake detection capability. Unlike the first two phases, which covered multiple facial features at once, this phase conducted separate tests for each facial region to gain detailed insights into how each specific area contributes to the detection process.

3.3. Model Selection

In the quest to detect DeepFakes through algorithmic means, we opted for two notable convolutional neural networks, EfficientNet-B7 and XceptionNet, serving as feature extractors. Both models utilise weights pre-trained on the ImageNet dataset, enabling them to leverage complex feature representations learned from a vast collection of natural images.

XceptionNet, conceived by François Chollet [34], is fundamentally characterised by its use of depth-wise separable convolutions. This innovative technique decomposes the standard convolution operation into depth-wise and point-wise convolutions, reducing the number of necessary parameters and computational burden. Consequently, this leads to faster training and improved model generalisation. Empirical evidence of XceptionNet's effectiveness in DeepFake detection is provided by Rossler et al. [31], validating its role as a primary feature extractor in our study.

In contrast, EfficientNet-B7 is the most advanced version in the EfficientNet series, known for its depth, parameter structure, and superior performance. It stands out by surpassing earlier models, achieving an accuracy of approximately 84.4% on the CIFAR-100 dataset and a top-5 accuracy of 97.3% on ImageNet [35]. Beyond these accuracy metrics, EfficientNet-B7 also offers significant benefits in model size and computational speed, being 8.4 times smaller and 6.1 times faster than the previous leading CNN model [36].

In summary, both EfficientNet-B7 and XceptionNet are chosen for their exceptional capabilities in handling complex image recognition tasks, such as DeepFake detection. EfficientNet-B7 is preferred for its superior performance, efficiency in scaling, and speed, while XceptionNet is valued for its innovative use of depth-wise separable convolutions, leading to efficient training and effective generalisation. Their pre-training on the ImageNet dataset further enhances their ability to act as powerful feature extractors in detecting DeepFakes.

3.4. Pre-Processing and Training Set-Up

This study utilised a dataset comprising authentic and manipulated videos in MP4 format. Key frames, specifically, every tenth, were extracted from each video for detailed analysis. The OpenCV library was crucial for isolating facial features by cropping these frames, while the MediaPipe library's face mesh function pinpointed precise facial landmarks as described in Section 3.1. These regions were subsequently obscured to focus on detecting facial augmentations, with the refined images organised into separate folders based on the excised facial areas.

All images in the datasets underwent a meticulous normalisation process, using channel-specific means of (0.485, 0.456, 0.406) and standard deviations of (0.229, 0.224, 0.225) for standardisation. Isotropic resizing was employed to unify image resolutions to 224 × 224 pixels, complemented by zero-padding for consistency. To enhance training data diversity, various augmentation techniques were implemented, including Image Compression, Gaussian Noise, and Flipping, each with a 10–15% probability of application. Notably, these augmentations were confined to the training dataset, preserving the integrity of the testing and validation datasets.

The training regimen leveraged the Rectified Adam optimiser [36], with an initial learning rate of 0.001 and a weight decay of 0.0005. A Reduction on Plateau strategy adjusted the learning rate, decreasing it by 0.25 when progress plateaued, with a patience parameter set at 2. The Binary Cross-Entropy Loss function was employed throughout the training, limited to 20 epochs as the maximum. An early stopping protocol was established to prevent overfitting, halting training if no improvement was observed over 10 consecutive epochs. The experiments consistently used a training batch size of 64, harnessing the power of an NVIDIA Geforce RTX 3080 Ti Laptop GPU and, for certain stages, Google Colab Pro+ was utilised for additional processing power.

Furthermore, EfficientNet-B7's pre-training involves a strategy termed Noisy Student, which introduces deliberate noise into the training data, enhancing the model's robustness to input variations and bolstering its performance in subsequent tasks. Due to these persuasive qualities, we incorporated EfficientNet-B7 as an additional feature extractor in our research.

3.5. Testing the Models

The model underwent evaluation by employing the pristine, non-augmented dataset, with a precise 10% allotment reserved explicitly for testing objectives. It warrants emphasis that this segregated test dataset was entirely distinct from the compilations utilised during the training and validation phases, a strategic measure to guarantee an assessment of the model's performance predicated on novel, previously unexposed data. For the purposes of this critical evaluation, only the facial regions encapsulated within the images were harnessed, as the crux of the performance appraisal was the model's proficiency in discriminating between authentic and fabricated faces, drawing upon the nuances of facial features as determinative criteria.

3.6. Libraries and Toolkits Used

This research heavily relied on various Python libraries for specialised tasks. OpenCV's VideoCapture was crucial for frame extraction from videos, while ImageDataGenerator was essential for data augmentation and preparation for training. Other libraries like Matplotlib, NumPy, and Pandas and random supported tasks such as visual representations, data manipulation, and randomisation were used.

For face detection and cropping, MediaPipe and its Face Mesh feature were used to identify and extract specific facial regions. Additionally, the scikit-learn library, known for its comprehensive machine learning algorithms and compatibility with libraries like NumPy and SciPy, was employed. This library facilitated the evaluation of models using metrics such as log-loss, vital for assessing performance and identifying areas for improvement, particularly in classification tasks.

4. Results

This research highlights the improved efficacy in DeepFake detection achieved through the strategic application of cutout techniques to distinct facial regions. By independently training and evaluating each specific area, the study conducts an in-depth examination of the contribution of various facial features to the accuracy of DeepFake identification. Two training subsets were established, with the first emphasising central facial features by covering the chin, mouth, jawline, and forehead, and the second concentrating on outer facial elements by covering the eyes and nose.

The research offers a comprehensive assessment of this technique's proficiency, incorporating visual analyses from face cutout augmentations derived from the FaceForensics++ and Celeb-DF datasets. It further includes a comparative analysis of models trained under three distinct scenarios:

- Baseline: Faces in their original, unaugmented form;
- Face-Cutout 1: Face cutout applied on peripheral facial features;
- Face-Cutout 2: Face cutout applied on the eyes and nose regions.

The results are juxtaposed with prevailing cutting-edge DeepFake detection methodologies, using identical datasets. The findings from each experimental segment are succinctly articulated, highlighting the significant revelations this method provides for the field of DeepFake detection.

4.1. Phase One: Evaluation of the Performance of the Cutout Technique with Each Dataset

During this phase, three image groups were generated from each dataset employed in the study: Baseline, Cutout 1, and Cutout 2. Subsequently, these groups were trained using the two deep convolutional models selected, which are EfficientNet-B7 and XceptionNet. Figure 5 provides a detailed assessment of the accuracy performance of the EfficientNet-B7 model during Phase One of our study. This figure specifically illustrates the outcomes of testing the model using two separate datasets: FF++ and Celeb-DF. The results indicate that the models trained with the Face-Cutout 2 group were significantly better compared to those for the Baseline and Face-Cutout 1 groups. Moreover, the results for the Cutout 1 group were worse than those for the Baseline group in some cases, as seen in the results of training with EfficientNet-B7 model, where the Cutout 1 group overperformed the Baseline group, as shown in Table 1.

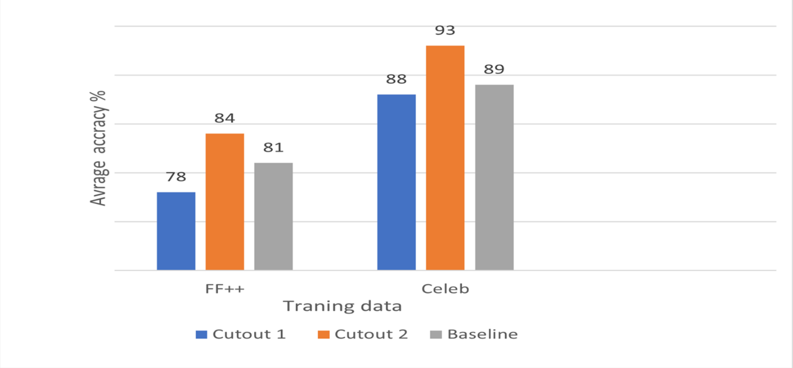


Figure 5. The accuracy performance of EfficientNet-B7 during Phase One.

Table 1. Phase One Results: Test results of the deep learning models trained separately using FaceForensics++ and Celeb-DF datasets.

Model	**FF++**			**Celeb-DF**		
	ACC	**AUC**	**Logloss**	**ACC**	**AUC**	**Logloss**
EfficientNet-B7 + Cutout 1	0.66	0.78	1.12	0.90	0.88	0.44
EfficientNet-B7+ Cutout 2	0.80	0.84	0.53	0.92	0.93	0.25
EfficientNet-B7+ Baseline	0.77	0.81	0.59	0.91	0.89	0.51
Xception + Cutout 1	0.75	0.81	0.90	0.90	0.91	0.35
Xception+ Cutout 2	0.75	0.83	0.76	0.91	0.92	0.29
Xception+ Baseline	0.77	0.77	0.78	0.84	0.79	0.80

Moreover, Table 1 shows the performance of two models, EfficientNet-B7 and XceptionNet, on the FF++ and Celeb datasets, trained for 20 epochs. The metrics used are AUC (area under the curve), ACC (accuracy), and log-loss. The Cutout data augmentation technique generally improved performance for both datasets, EfficientNet-B7 and XceptionNet.

The results demonstrate significant improvements in the performance of the EfficientNet and Xception models for the Cutout 2 group, with accuracy gains varying between 1.23% and 17.7%, compared to the Baseline group. These findings highlight the effectiveness of the methods implemented with the Cutout 2 dataset in developing more resilient facial recognition models. This improvement can be ascribed to the targeted learning approach within the Cutout 2 dataset, which necessitates that models distinguish faces by concentrating on the most informative facial areas. This strategy is based on the premise that these regions provide critical cues for differentiating between fake and real faces.

In the context of the Celeb-DF dataset, the Cutout 2 group log-loss results improved by 43.18% compared to the Cutout 1 group when using the EfficientNet-B7 model, and this was nearly matched by the performance of the Xception model. The models trained with Face-Cutout 2 augmentations clearly exhibited superior performance compared to the other two conditions. Consequently, training was also applied with the Forensics++ dataset.

In training deep neural models with Face-Cutout 2 images, the models emphasised the exposed facial regions. In the Cutout 2 group, the obscured areas included the left eye, right eye, both eyes, or the nose. From these results, we infer that the face regions outside these specified areas offer more critical information for distinguishing between authentic and synthetic faces.

As a result, the superior performance of the Cutout 2 augmentations is attributed to the inclusion of facial features beyond the eyes and nose, which are typically central to facial recognition. This approach enhances the reliability of detecting facial differences.

Huang et al. [37] explored expression recognition with partially occluded faces and discovered that their model could still identify most facial expressions, even with the eyes occluded, by relying on the face's external features. This aligns with the findings of our current study, which suggest that facial features outside the central area hold substantial information that is crucial for distinguishing between similar faces. Therefore, it can be concluded that, in scenarios where faces are strikingly similar, as with DeepFakes, detecting differences between them is more accurate when the focus is on facial features beyond the central region.

4.2. Phase Two: Evaluation of the Performance of Cutout Technique with the Combined Dataset

In this phase, the datasets used in Phase One were combined to increase the overall volume of training data, which could potentially improve the model's generalisation and performance with previously unseen data. Furthermore, the model will have more diverse examples to learn from, which can help it to better understand underlying patterns and features in the data. Table 2 shows the results of the second phase of the experiment, which evaluated the performance of the three different groups (Baseline, Cutout 1, and Cutout 2) with a combined dataset of face images. The models were trained for 20 epochs and their performance was evaluated in terms of AUC, ACC (accuracy), and log-loss metrics. The results show that the EfficientNet-B7 model with Cutout 2 achieved the best performance, with an AUC of 0.89, ACC of 0.91, and log-loss of 0.45. The Xception model with Cutout 2 also performed well, with an AUC of 0.86, ACC of 0.88, and log-loss of 0.85. The EfficientNet-B7 and the Xception models with Baseline datasets performed similarly, with AUC values of 0.90 and 0.83 respectively.

Table 2. Phase two results with combined datasets.

Model/Cutout Type	**Combined Dataset**		
	ACC	**AUC**	**Logloss**
EfficientNet-B7 + Cutout 1	0.89	0.90	0.48
EfficientNet-B7 + Cutout 2	0.89	0.91	0.45
EfficientNet-B7 + Baseline	0.90	0.87	0.69
Xception + Cutout 1	0.83	0.85	0.84
Xception + Cutout 2	0.86	0.88	0.85
Xception + Baseline	0.77	0.73	0.94

The findings of this experiment indicate that employing the Cutout 2 technique enhances the performance of face recognition models more effectively than the Cutout 1 and Baseline methods. Additionally, the results point to the EfficientNet-B7 model being superior in face recognition compared to the Xception model.

In the Cutout 2 group, the generated cutouts seem to preserve facial features more effectively. This likely assists the model in learning these features more efficiently, thereby boosting face-recognition performance. Overall, according to the data presented in Table 2, the Cutout 2 method emerges as a promising approach for DeepFake identification.

The AI models were more effective in identifying DeepFakes when they utilised external regions of the face, such as the forehead, cheeks, and chin. This indicates that these external regions might contain more valuable information for distinguishing between real and fake faces.

Overall, the findings imply that the external facial regions could be as crucial, or perhaps even more critical, than the core regions of the face in detecting DeepFakes. This is a significant discovery, as it suggests that AI models can identify DeepFakes even when the eyes are not visible.

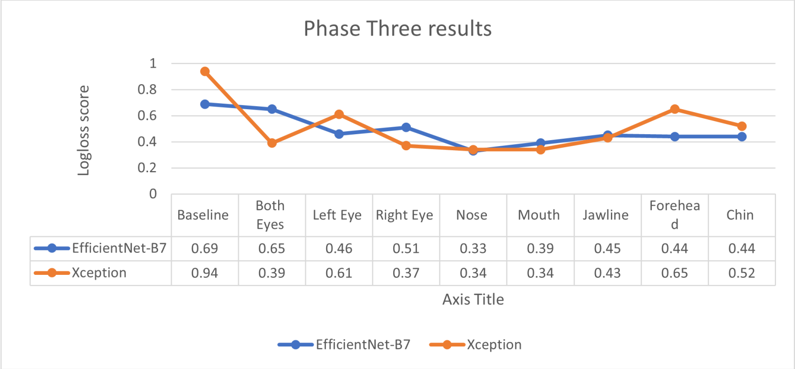
To provide a more detailed understanding of the contribution of each facial region to DeepFake detection, separate tests were conducted for each region. These are explained in greater detail in the following section.

4.3. Phase Three: Evaluation of the Cutout Technique for Each Facial Feature

To gain a fuller understanding of the facial attributes that contribute substantial information in the detection of DeepFake images, a further, more detailed investigation was undertaken. This involved the application of cutout techniques on distinct facial regions, thereby facilitating independent training and evaluation for each region. Multiple training groups were generated, each of which comprised images exclusively featuring a single face cutout feature, such as the eyes or mouth. A comprehensive evaluation was conducted to assess the efficacy of each training group in allowing the successful discrimination between synthetic and authentic images. The results are summarised in Figure 6. This figure primarily showcases the performance of EfficientNet-B7 and Xception models, which were individually trained on eight distinct face cutouts. The evaluation of these cutouts, carried out with great precision, forms a critical part of the analysis in Phase Three of our research. The following findings are notable:

- EfficientNet-B7 has a lower log-loss score than Xception for all facial features except the left eye. This means that EfficientNet-B7 is better at classifying facial features than Xception. The difference in performance is most pronounced for the nose, mouth, and jawline;
- The Baseline log-loss score is very high, which means that the group where no facial cutout is used is not very good at identifying DeepFakes. However, the use of either or both eyes, the nose, mouth, jawline, forehead, and chin all help to improve the accuracy of facial feature classification;
- The log-loss score for the right eye cutout is similar to that for the left eye. This suggests that the right eye is just as important as the left eye for DeepFake classification;
- Among the cutouts used for image classification, the nose is the most critical feature. The log-loss score for the nose cutout is much lower than those for the eyes, and a lower log-loss value indicates that the model is more likely to correctly classify the image. The fact that the model can still perform well without using information about the nose suggests that the nose region provides less information for the identification of DeepFakes than the eyes;
- The mouth cutout is almost as important as the nose for DeepFake classification. This is because the log-loss score for the mouth is similar to that for the nose. Based on these results, it can be concluded that the nose and mouth regions provide less information for the classification of DeepFakes than the eyes;
- The jawline cutout is slightly less important than the mouth for facial feature classification. This is because its log-loss score is slightly higher than that for the mouth;
- Of all the facial features used to classify DeepFakes, the forehead, both eyes, and chin have the highest log-loss values, as shown in Figure 6. This means that the model is not very good at classifying DeepFakes when these regions are blocked out. This suggests that the forehead, eyes, and chin contain significant information for the classification of DeepFakes.

These findings are in line with those of Das et al. [13], who also trained the EfficientNet-B4 and Xception models with three groups of images with the facial features of the eyes, nose, or mouth obscured. They found that the model trained on images with obscured noses and mouths performed best, while the model trained on images with obscured eyes performed worst. These consistent outcomes between the present study and the findings of Das et al. [13] strengthen the validity and reliability of the observed trends concerning the impact of facial region cutouts on model performance in classifying DeepFakes.



	Baseline	Both Eyes	Left Eye	Right Eye	Nose	Mouth	Jawline	Forehea d	Chin
EfficientNet-B7	0.69	0.65	0.46	0.51	0.33	0.39	0.45	0.44	0.44
Xception	0.94	0.39	0.61	0.37	0.34	0.34	0.43	0.65	0.52

Figure 6. Test results of the deep learning models trained separately using the datasets generated in Phase Three.

4.4. *Comparison with State-of-the-Art Methods*

Given the markedly better performance exhibited by the Face-Cutout 2 group when compared to the Cutout 1 and Baseline groups, a thorough comparative analysis was conducted to compare the approach used in the present study with state-of-the-art methodologies that have employed identical datasets and techniques. This evaluation aimed to assess the effectiveness and robustness of the proposed technique in relation to existing approaches within the same experimental setting. The aim of this rigorous comparative analysis was to establish the competitiveness and potential advantages of the proposed method in the realm of DeepFake detection.

4.4.1. DeepFake Detection Approaches Using the FF++ Dataset

Das et al. [13] avoided using random cutouts of the face and instead selected specific regions to be covered in each group. Their study employed the removal of data in two groups: the 'sensory' and 'convex-hull' groups. The former group achieved the best performance and covered the eyes, nose, and mouth regions. In fact, this group outperformed the Face-Cutout 2 group in the present study. The reason for this may be that Das et al.'s sensory group covered three facial regions, two of which (the nose and mouth) were found to be less important for DeepFake detection compared to other regions of the face, as shown in Table 3. Moreover, Table 3 provides a detailed comparative evaluation of the accuracy performance of several baseline models in the context of DeepFake detection. This comparison specifically focuses on the results obtained using the FaceForensics++ dataset, illustrating how each model fares in identifying DeepFake content within this particular data environment. Moreover, the sensory group utilised both of these less-important regions, which allowed the model to benefit from the regions of the face that provided more information than others.

Table 3. A comparison of the performance (accuracy) of various baseline models for detecting DeepFakes on the FaceForensics++ dataset.

Approach	Accuracy	AUC	Number of Datasets	Year
Das et al. [13]	N/A	96.73	N/A	2021
Khan and Dang-Nguyen [14]	95.57	N/A	200,000	2022
Lee et al. [30] face-patch	72.79	N/A	60,000	2022
Lee et al. [30] face-crop	80.56	N/A	60,000	2022
Rossler et al. [31]	90.60	N/A	388,000	2019
Afchar et al. [38]	83.10	N/A	16,000	2018
Zhang et al. [39]	79.09	72.22	6706	2021
Our Face-Cutout	84	80	10,000	2023

In contrast, the methodology employed in the current study adheres to specific guidelines throughout the face cutout procedure. For instance, in the Face-Cutout 2 group, the exclusion of landmarks focused on specific regions: the left eye, right eye, both eyes, or nose. This purposeful selection of regions allowed the models to give preference to facial features and focus their attention on information extracted from the external facets of the face. Compared to other methodologies, as shown in Table 4, the models in this study were trained with a smaller sample size compared to other methodologies. For instance, Rossler et al. [31] conducted training on a dataset consisting of approximately 388,000 images, while Khan and Dang-Nguyen [14] employed around 200,000 images when training their models. This divergence in the size of the training dataset may help explain the superior accuracy attained by these studies in comparison to the approach used here.

Nevertheless, our study exhibited superior performance compared to that by Lee et al. [30] for the face-patch group. Their research achieved an accuracy rate of 72.79% by covering the mouth and nose. In contrast, our approach involved covering the eyes and nose and, notably, we worked with a smaller dataset than that employed in their study. In addition, the model in the present study outperformed those in other studies such as Afchar et al. [38] and Zhang et al. [39], which utilised the FF++ dataset for DeepFake detection. Meanwhile, although Zhang et al. achieved an accuracy of 79.09% by randomly dropping out parts of the frames, in the present study, an accuracy level of 84% was achieved. This may be said to confirm the hypothesis that the identification of facial differences is facilitated by the utilisation of facial features that are not exclusively located at the centre of the face.

Another factor influencing the outcome is the approach used to select facial landmarks and to conduct the cutout procedure. In both of the aforementioned studies, landmarks to be covered in specific facial regions in the images were selected at random, resulting in the model being exposed to all areas of the face within the same group of data. Consequently, there were no constraints preventing certain parts of the face from appearing in specific subsets of the dataset, making it more difficult to determine which facial components exerted more influence than others. The novel facial Cutout 2 could reduce overfitting and enhance the model's detection capabilities. Additionally, it is also demonstrated that the model used in the present study can learn from a smaller volume of data.

4.4.2. DeepFake Detection Approaches Using the Celeb-DF Dataset

The Celeb-DF dataset is a widely used benchmark dataset used in the evaluation of the performance of DeepFake detection models, since it contains a large number of high-quality videos of celebrities that have been manipulated to create DeepFakes. By comparing the performance of various baseline models with this dataset, insights can be gained into which models are most effective in detecting DeepFakes and how they compare with each other in terms of accuracy. Table 4 showcases a comprehensive comparison among different baseline models utilised for DeepFake detection, specifically focusing on their application to the Celeb-DF dataset. This table highlights the peak performance metrics achieved by each model, providing a detailed insight into their effectiveness and efficiency in accurately identifying DeepFakes within the Celeb-DF dataset environment.

Table 4. Comparison of the best results achieved by DeepFake detection baseline models when used with the Celeb-DF dataset.

Approaches	Accuracy	AUC	Year
Haliassos et al. [40]	82.4	0.85	2021
Zhang et al. [39]	81.08	88	2021
Ismail et al. [41]	90.73	90.62	2021
Li, W., and Shen, Z. [42]	83.81	–	2022
Lee et al. [15]	–	76.50	2023
Li et al. [43]	80.58	84	2020
Masi et al. [44]	76.6	82	2020
Present study, Face-Cutout 2	**92**	**93**	2023

It can be observed that the approach proposed in this study achieves an impressive level of performance that is higher than that of all of the state-of-the-art models, with an accuracy of 92% and AUC of 93%, thus demonstrating the effectiveness and ability of this approach in handling various DeepFake generation methods.

Comparing the approach used in the present study to that of Zhang et al. [39], it is observed that their method, which involved randomly dropping out parts of the frames, achieved an impressive AUC of 88.83%. This is a good performance level compared to the present approach, which achieved an AUC of 93%. Compared to our face Cut-out 2 method, the approach in [41], using the YOLO face detector and InceptionResNetV2 CNN, presents a different strategy in DeepFake detection. In their method, the YOLO detector extracts the facial area from video frames, and the InceptionResNetV2 CNN is then used to extract features from these faces. These extracted features are subsequently fed into an XGBoost classifier that operates at the top level of the CNN network. This method has achieved an impressive 90.62% AUC and 90.73% accuracy.

In our study , the "Face Cut-out 2" method showed a notable improvement in performance, achieving a 93% AUC on the Celeb-DF dataset. This markedly surpasses the highest accuracy of 83.51% recorded in the FD2Foremer study [42] using their Img+ detail (swin) technique. Their method focuses on examining mid-frequency facial geometry details, encompassing individual-specific characteristics as well as dynamic, expression-related features on the face.

Another study [15] introduces an alternative innovative method for DeepFake detection, which utilises a patch-by-patch No-Reference Image Quality Assessment to distinguish between facial and non-facial regions. This is complemented by a frequency-decomposition block that extracts both high- and low-frequency components. While this technique achieved a respectable AUC of 76.50 on the Celeb dataset and showed promise, it does not quite match the performance we achieved with our method on the same dataset. This indicates that our model is capable of developing more robust representations compared to this earlier approach.

As a concluding point in this comparison, it is important to highlight that our study adopted a unique approach by selectively obscuring different parts of the face. This method shed light on the critical role various facial features play in DeepFake detection. The insights gained from this approach are instrumental in guiding future research efforts to improve the precision and reliability of DeepFake detection models.

4.4.3. DeepFake Detection Approaches that Used Similar Techniques

Deep neural network models are widely utilised for detection purposes, with the choice of specific algorithms varying based on data type and the insights gained during training. For instance, the FDML model in a study [45] focuses on detecting fake news, especially in brief content, by combining fake news detection with news topic classification. It employs a unique news graph method and dynamic weighting strategy.

Contrastingly, the "FraudTrip" study [46] deals with identifying fraudulent taxi trips by analysing GPS data, diverging from textual content analysis. In our research, we concentrate on facial image analysis to distinguish DeepFake images, using a technique called "face cut-out" to analyse the importance of different facial regions in DeepFake detection. These diverse applications highlight the adaptability of neural networks in handling various data types and objectives, from multimedia content in fake news to spatial data in fraud detection and facial feature analysis in DeepFake identification.

To further emphasise the distinction and superiority of our approach in the realm of DeepFake detection, the comparative analysis presented in Table 5 plays a pivotal role. This table meticulously outlines the performance metrics of various DeepFake detection methods that share a resemblance to our occlusion-based technique. It offers an exhaustive comparison, underscoring the efficacy of each method, particularly in their ability to accurately identify DeepFakes. This comparison is not just a measure of performance but a critical assessment of how occlusion-based strategies fare against traditional DeepFake detection methods, thereby enriching our understanding of their place and potency in the field.

Our model's standout performance, with an accuracy of 91%, is not just a numerical lead but a testament to the innovative approach we have adopted. The detailed breakdown in the second column of the table, which illustrates the specific areas of the images that were obscured during training, reveals a significant insight: the precision of occlusion matters. Methods that employed random cutouts, though innovative, lacked the targeted effectiveness, resulting in lower accuracy. Our approach, conversely, strategically selects specific facial regions for occlusion, a method that has proven to be more than just a novel technique but a highly effective one at that.

This nuanced approach to occlusion, focusing on specific facial regions, does more than just improve accuracy; it opens up new avenues for DeepFake detection. This suggests a path forward where the focus is not just on detecting falsities but on understanding the dynamics of image manipulation. The superior performance of our method indicates a potential shift in the paradigm of DeepFake detection—from broad-stroke analysis to more refined, detail-oriented strategies.

Table 5. Results for DeepFake detection state-of-the-art techniques similar to that used in this study.

The Method	The Occlusion Part	Acc
[28]	Randomly erasing a rectangular region in the image	76.7
[37]	Occlusion approach integrating features of eyes, nose, and mouth	87.08
[47]	Utilised varied objects like sunglasses and masks for facial occlusion	77.4
[48]	Random square cutout in the image	81.0
Present study	Cut out one of four regions: left/right eye, both eyes, or nose	**91**

In conclusion, the findings from this comparative analysis do not merely place our method at the forefront of DeepFake detection; they potentially revolutionise the approach towards more sophisticated, reliable, and effective DeepFake detection methodologies. This could have far-reaching implications, not just in enhancing current technologies but in shaping future research and development in the field.

Our method shows strong potential for real-time DeepFake detection, primarily due to its advanced feature recognition and analysis within a GAN-based model. It excels in identifying DeepFakes by focusing on specific facial features, even in challenging scenarios like occlusions from sunglasses or masks. This strategy [49] is reminiscent of techniques used in underwater image analysis, where a focus on certain background elements yielded improved results.

However, one limitation is that our model primarily recognises front-facing features of the face, so different angles could reduce effectiveness. Moreover, the computational intensity required for these sophisticated algorithms, including GANs, poses a challenge for real-time application due to the high demand for processing power. Thus, while promising, our method needs refinement in handling diverse facial orientations and computational efficiency for effective real-time DeepFake detection.

5. Conclusions

In this study, we compared the performance of two deep learning models for the classification of DeepFakes and investigated the importance of different facial features for identification of DeepFakes. The results suggest that EfficientNet-B7 is a better model for the classification of DeepFakes than Xception. Our findings present an interesting dynamic in how different facial features influence the performance of the DeepFake detection model in two scenarios:

- Covering Multiple Facial Features: When we covered multiple facial regions at once, the Face-Cutout 2 group, where internal features (like the eyes, nose, and mouth) were obscured, showed higher accuracy (up to 91%) compared to the Face-Cutout 1 group that achieved 88% accuracy with external features (like hairline, ears, and jawline) covered. This suggests that, overall, the model may rely more heavily on external features for DeepFake detection when multiple areas are obscured;

- Covering Individual Facial Features: In individual tests of facial features, a distinct trend emerged. The model's performance was most adversely affected when the eyes were covered, leading to a performance decrease, with an average log-loss score of 0.52. In contrast, covering the nose yielded the most favorable results, with a log-loss score of 0.33.

This research introduces a novel method: using Face-Cutout techniques to better understand key facial features that are crucial in DeepFake detection. This approach allows for a detailed analysis of facial regions, significantly advancing DeepFake detection by pinpointing the most informative facial elements. Furthermore, this study introduces a novel method that uses Face-Cutout techniques to better understand key facial features that are crucial in DeepFake detection. This approach allows for a detailed analysis of facial regions, significantly advancing DeepFake detection by pinpointing the most informative facial elements.

Our study faced key limitations, including difficulty in obtaining diverse, high-quality training data with both real and manipulated samples. Time constraints were significant, as dataset preparation and training, often requiring manual effort, were time-intensive. Additionally, the nascent field of DeepFake detection suffers from unclear methodologies and limited comprehensive research, complicating result replication and understanding.

Future research should explore additional deep learning models, like ResNet and VGGNet, for DeepFake classification, beyond EfficientNet-B7 and Xception used in this study. It is also important to assess the proposed approach with larger datasets and to extend the study to video-based DeepFake detection methods, which utilise spatial and temporal data from consecutive frames.

Author Contributions: Conceptualisation, F.A.; methodology, F.A.; validation, F.A.; formal analysis, F.A.; investigation, F.A. and G.M.; resources, F.A.; writing—original draft preparation, F.A.; writing—review and editing, G.U.; visualisation, F.A.; supervision, G.M. and G.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Dataset utilized in this research is publicly available: https://github.com/yuezunli/celeb-deepfakeforensics (accessed on 15 June 2023) and https://github.com/ondyari/FaceForensics (accessed on 18 June 2023).

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

References

1. Masood, M.; Nawaz, M.; Malik, K.M.; Javed, A.; Irtaza, A.; Malik, H. Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way Forward. *Appl. Intell.* **2023**, *53*, 3974–4026. [CrossRef]
2. Vasist, P.N.; Krishnan, S. Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research. *Commun. Assoc. Inf. Syst.* **2022**, *51*, 14.
3. Chen, J.; Wang, Q.; Peng, W.; Xu, H.; Li, X.; Xu, W. Disparity-based Multiscale Fusion Network for Transportation Detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 18855–18863. [CrossRef]
4. Xu, H.; Han, S.; Li, X.; Han, Z. Anomaly Traffic Detection Based on Communication-Efficient Federated Learning in Space-Air-Ground Integration Network. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 9346–9360. [CrossRef]
5. Dong, W.; Yang, Y.; Qu, J.; Xiao, S.; Li, Y. Local Information-Enhanced Graph-Transformer for Hyperspectral Image Change Detection With Limited Training Samples. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5509814. [CrossRef]
6. Yan, L.; Shi, Y.; Wei, M.; Wu, Y. Multi-Feature Fusing Local Directional Ternary Pattern for Facial Expressions Signal Recognition Based on Video Communication System. *Alex. Eng. J.* **2023**, *63*, 307–320. [CrossRef]
7. Tao, Y.; Shi, J.; Guo, W.; Zheng, J. Convolutional Neural Network Based Defect Recognition Model for Phased Array Ultrasonic Testing Images of Electrofusion Joints. *J. Press. Vessel Technol.* **2023**, *145*, 024502. [CrossRef]
8. Jannat, M.K.A.; Islam, M.S.; Yang, S.; Liu, H. Efficient Wi-Fi-Based Human Activity Recognition Using Adaptive Antenna Elimination. *IEEE Access* **2023**, *11*, 105440–105454. [CrossRef]
9. Westerlund, M. The Emergence of Deepfake Technology: A Review. *Technol. Innov. Manag. Rev.* **2019**, *9*, 39–52 . [CrossRef]

10. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
11. Bitouk, D.; Kumar, N.; Dhillon, S.; Belhumeur, P.; Nayar, S.K. Face Swapping: Automatically Replacing Faces in Photographs. *ACM Trans. Graph.* **2008**, *27*, 1–8. [CrossRef]
12. Korshunova, I.; Shi, W.; Dambre, J.; Theis, L. Fast Face-Swap Using Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3677–3685.
13. Das, S.; Seferbekov, S.; Datta, A.; Islam, M.S.; Amin, M.R. Towards Solving the Deepfake Problem: An Analysis on Improving Deepfake Detection Using Dynamic Face Augmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3776–3785.
14. Khan, S.A.; Dang-Nguyen, D.T. Hybrid Transformer Network for Deepfake Detection. In Proceedings of the 19th International Conference on Content-based Multimedia Indexing, Graz, Austria, 14–16 September 2022; pp. 8–14.
15. Lee, E.G.; Lee, I.; Yoo, S.B. ClueCatcher: Catching Domain-Wise Independent Clues for Deepfake Detection. *Mathematics* **2023**, *11*, 3952. [CrossRef]
16. Tolosana, R.; Romero-Tapiador, S.; Fierrez, J.; Vera-Rodriguez, R. Deepfakes Evolution: Analysis of Facial Regions and Fake Detection Performance. In *International Conference on Pattern Recognition*; Springer International Publishing: Cham, Switzerland, 2021; pp. 442–456.
17. Waseem, S.; Abu-Bakar, S.A.R.S.; Omar, Z.; Ahmed, B.A.; Baloch, S.; Hafeezallah, A. Multi-Attention-Based Approach for Deepfake Face and Expression Swap Detection and Localization. *EURASIP J. Image Video Process.* **2023**, *1*, 14. [CrossRef]
18. Silva, S.H.; Bethany, M.; Votto, A.M.; Scarff, I.H.; Beebe, N.; Najafirad, P. Deepfake Forensics Analysis: An Explainable Hierarchical Ensemble of Weakly Supervised Models. *Forensic Sci. Int. Synerg.* **2022**, *4*, 100217. [CrossRef] [PubMed]
19. Le, B.; Tariq, S.; Abuadbba, A.; Moore, K.; Woo, S. Why Do Facial Deepfake Detectors Fail? In Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes, Melbourne, Australia, 10–14 July 2023; pp. 24–28.
20. Mahmud, B.U.; Sharmin, A. Deep Insights of Deepfake Technology: A Review. *arXiv* **2021**, arXiv:2105.00192.
21. Shahzad, H.F.; Rustam, F.; Flores, E.S.; Mazón, J.L.V.; Diez, I.d.l.T.; Ashraf, I. A Review of Image Processing Techniques for Deepfakes. *Sensors* **2022**, *22*, 4556. [CrossRef]
22. Malik, A.; Kuribayashi, M.; Abdullahi, S.M.; Khan, A.N. DeepFake Detection for Human Face Images and Videos: A Survey. *IEEE Access* **2022**, *10*, 18757–18775. [CrossRef]
23. Jung, T.; Kim, S.; Kim, K. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access* **2020**, *8*, 83144–83154. [CrossRef]
24. Menotti, D.; Chiachia, G.; Pinto, A.; Schwartz, W.R.; Pedrini, H.; Falcao, A.X.; Rocha, A. Deep Representations for Iris, Face, and Fingerprint Spoofing Detection. *IEEE Trans. Inf. Forensics Secur.* **2015**, *10*, 864–879. [CrossRef]
25. Ciftci, U.A.; Demir, I.; Yin, L. Fakecatcher: Detection of Synthetic Portrait Videos Using Biological Signals. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, 1–17. [CrossRef]
26. Jafar, M.T.; Ababneh, M.; Al-Zoube, M.; Elhassan, A. Forensics and Analysis of Deepfake Videos. In Proceedings of the 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 7–9 April 2020; pp. 53–58.
27. Ge, S.; Li, J.; Ye, Q.; Luo, Z. Detecting Masked Faces in the Wild with LLE-CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
28. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13001–13008. [CrossRef]
29. Chen, L.; Zhang, Y.; Song, Y.; Liu, L.; Wang, J. Self-Supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18710–18719.
30. Ko, D.; Lee, S.; Park, J.; Shin, S.; Hong, D.; Woo, S.S. Deepfake Detection for Facial Images with Facemasks. *arXiv* **2022**, arXiv:2202.11359.
31. 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, Republic of Korea, 27 October–2 November 2019; pp. 1–11.
32. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3207–3216.
33. Google. Mediapipe Face Mesh Documentation. 2022. Available online: https://github.com/google/mediapipe/blob/master/docs/solutions/face_mesh.md (accessed on 6 May 2023).
34. 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.
35. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
36. Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Han, J. On the Variance of the Adaptive Learning Rate and Beyond. *arXiv* **2019**, arXiv:1908.03265.
37. Huang, X.; Zhao, G.; Zheng, W.; Pietikäinen, M. Towards a Dynamic Expression Recognition System Under Facial Occlusion. *Pattern Recognit. Lett.* **2012**, *33*, 2181–2191. [CrossRef]

38. 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.
39. Zhang, D.; Li, C.; Lin, F.; Zeng, D.; Ge, S. Detecting Deepfake Videos with Temporal Dropout 3DCNN. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Montreal, QC, Canada, 19–27 August 2021; pp. 1288–1294.
40. 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, 20–25 June 2021; pp. 5039–5049.
41. Ismail, A.; Elpeltagy, M.; Zaki, M.S.; Eldahshan, K. A new deep learning-based methodology for video deepfake detection using xgboost. *Sensors* **2021**, *21*, 5413. [CrossRef]
42. Li, W.; Shen, Z. FD 2 Foremer: Thinking Face Forgery Detection in Midfrequency Geometry Details. *Secur. Commun. Netw.* **2022**, *2022*, 9278715. [CrossRef]
43. 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, 13–19 June 2020; pp. 5001–5010.
44. 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 16th European Conference on Computer Vision (ECCV 2020), Glasgow, UK, 23–28 August 2020; Part VII 16 , pp. 667–684.
45. Liao, Q.; Chai, H.; Han, H.; Zhang, X.; Wang, X.; Xia, W.; Ding, Y. An Integrated Multi-Task Model for Fake News Detection. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5154–5165. [CrossRef]
46. Ding, Y.; Zhang, W.; Zhou, X.; Liao, Q.; Luo, Q.; Ni, L.M. FraudTrip: Taxi Fraudulent Trip Detection from Corresponding Trajectories. *IEEE Internet Things J.* **2020**, *8*, 12505–12517. [CrossRef]
47. Han, J. Face Analysis and Deepfake Detection. Ph.D. Thesis , University of Amsterdam, Amsterdam, The Netherlands, 2021.
48. DeVries, T.; Taylor, G.W. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv* **2017**, arXiv:1708.04552.
49. Yang, M.; Wang, H.; Hu, K.; Yin, G.; Wei, Z. IA-Net: An Inception–Attention-Module-Based Network for Classifying Underwater Images from Others. *IEEE J. Ocean. Eng.* **2022**, *47*, 704–717. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.