

XDF: A Large-Scale Dataset for Evaluating Video Deepfake Detection Across Multiple Manipulation Techniques

Vazgken Vanian

vvanian@iti.gr

Centre for Research and Technology Hellas

Thessaloniki, Greece

Konstantinos Konstantoudakis

k.konstantoudakis@iti.gr

Centre for Research and Technology Hellas

Thessaloniki, Greece

Georgios Petmezas

petmezgs@iti.gr

Centre for Research and Technology Hellas

Thessaloniki, Greece

Dimitris Zarpalas

zarpalas@iti.gr

Centre for Research and Technology Hellas

Thessaloniki, Greece

Abstract

Deepfake technologies have rapidly advanced, presenting significant challenges to the integrity of digital media and creating potential risks in various sectors, from politics to personal privacy. In response, the research community has focused on developing reliable deepfake detection methods. However, the continuous advancements and growing complexity of artificial intelligence (AI) models have outpaced existing datasets, making it difficult to train and evaluate detection systems effectively. This paper addresses this gap by introducing a comprehensive dataset of real and manipulated videos, aimed at supporting the development and evaluation of advanced deepfake detection models. This dataset could serve as a benchmark for assessing the effectiveness of emerging detection methodologies, providing a standardized resource for researchers to measure progress and compare results. Additionally, this study offers key insights for the creation of effective deepfake datasets and identifies several pressing challenges in the field, guiding future research and development efforts.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence; Computer vision.**

Keywords

Deepfake Detection, Deepfake Dataset, Face-Swapping, Facial-Reenactment, Lip-Syncing

ACM Reference Format:

Vazgken Vanian, Georgios Petmezas, Konstantinos Konstantoudakis, and Dimitris Zarpalas. 2025. XDF: A Large-Scale Dataset for Evaluating Video Deepfake Detection Across Multiple Manipulation Techniques. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25)*, June 30–July 3, 2025, Chicago, USA (MAD2025). ACM, New York, NY, USA, 8 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MAD2025, Chicago, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

The rise of AI has dramatically transformed digital content creation giving rise to deepfake technology, which generates highly convincing manipulated videos [10]. Deepfakes can alter or replace a person's likeness in a video, making it appear as though they said or did things they never actually did. While deepfake technology holds promise for creative applications, its potential misuse has also raised significant concerns, especially regarding identity theft, misinformation and security threats [15].

In the early stages, deepfakes were relatively easy to detect, often exhibiting noticeable imperfections, such as unnatural facial expressions or inconsistent lighting. However, with the rapid progress of deep learning (DL) techniques, the quality and realism of deepfakes have improved dramatically, making them increasingly difficult to distinguish from authentic footage. These advances have made it all the more urgent to develop robust detection methods to counter the potential threats posed by these manipulations [17].

To address the escalating risks associated with deepfakes, researchers have developed a wide range of detection methods, primarily leveraging AI-driven approaches such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [19]. These models excel at analyzing both spatial and temporal patterns in videos, identifying subtle artifacts or inconsistencies that may not be visible to the human eye. By focusing on minute irregularities in facial movements, lighting or texture inconsistencies, these models have demonstrated impressive performance in detecting manipulated content.

However, despite notable progress, many existing detection systems face significant limitations [16]. A key challenge is their tendency to overfit to specific datasets or manipulation techniques, resulting in reduced generalization to unseen data or deepfakes generated using novel methods. Furthermore, the high computational cost of training and running these models can hinder their practical deployment in real-time applications. Another critical issue is the scarcity of diverse, open-access deepfake datasets that can be used to develop and benchmark detection methods [7]. Most publicly available datasets are limited in terms of manipulation techniques, video quality and diversity of subjects, which further restricts the robustness of detection models.

This study aims to bridge this gap by introducing XDF (eXtended DeepFake) dataset, a large-scale, open-access video deepfake dataset that provides a comprehensive set of both original and manipulated

videos across a wide range of techniques. Our dataset contains in total 80,000 video sequences, making it one of the largest and most diverse deepfake datasets available. These include 20,000 authentic sequences alongside 60,000 deepfake sequences generated using six different deepfake creation tools that correspond to three distinct manipulation techniques, namely face-swapping, facial-reenactment and lip-syncing.

XDF includes both original and manipulated sequences of the same individuals, offering a unique resource for developing both binary classification models and approaches that leverage identity verification. Additionally, it encompasses a diverse range of real-world conditions, such as varying backgrounds, lighting and camera angles, making it a valuable benchmark for evaluating the robustness and generalizability of video deepfake detection methods. Comparisons with other highly cited deepfake datasets demonstrate that incorporating XDF in future research is crucial for enhancing model performance and ensuring generalization across various manipulation techniques.

Our contributions can be summarized as follows:

- We present a novel deepfake dataset that incorporates a variety of manipulation techniques, offering enhanced robustness for detecting different forms of deepfakes.
- We evaluate the performance of the proposed dataset compared to existing ones using state-of-the-art deepfake detection models.
- We demonstrate that incorporating our dataset as part of the training set significantly improves model generalization to other datasets, addressing a major challenge in deepfake detection by enhancing cross-dataset performance.
- We provide valuable insights and recommendations for future work in deepfake detection including key considerations for creating comprehensive and effective datasets, guiding further research and dataset development in the field.

2 Related Work

2.1 Video Deepfake Datasets

A wide array of datasets has been created to facilitate the development of video deepfake detection techniques. Each dataset has contributed significantly to the field, but limitations such as size, diversity and manipulation techniques still hinder comprehensive model generalization.

One of the earliest datasets, UADFV [27], introduced a small set of 49 real and 49 fake videos generated using a basic face-swapping method. Although it provided a foundation for early research, its limited scale and narrow focus on a single manipulation technique make it insufficient for modern detection challenges. Similarly, Deepfake-TIMIT [11] featured 640 manipulated videos using two face-swapping techniques, but the videos were synthetic and controlled, lacking real-world variability in lighting, background or camera angles.

FaceForensics++ (FF++) [21], one of the most influential datasets, significantly expanded the available data with 1,000 real videos and 5,000 manipulated ones, created using multiple techniques including face-swapping and facial-reenactment. This dataset introduced compression artifacts to simulate real-world conditions, but is still somewhat limited in the range of manipulation tools and identities it

covers. Celeb-DF [13], another key resource, provided 5,639 deepfake videos based on celebrity footage. While Celeb-DF focused on high-quality face-swaps, its smaller pool of identities can limit the generalization of models trained on it.

The DeepFake Detection Challenge (DFDC) dataset [5], sponsored by Facebook, marked a significant leap forward with over 100,000 videos generated using multiple methods. However, the manipulations are predominantly face-swaps and the dataset does not provide the same controlled authentic versus manipulated comparisons for each individual that are critical for binary classification.

Finally, DeeperForensics-1.0 [9] addresses some of the limitations found in earlier datasets by offering a large-scale set of 50,000 manipulated videos and 10,000 authentic ones. These videos were created using high-quality face-swapping techniques and involve various real-world scenarios, including different lighting conditions, camera angles and compression artifacts, making it more representative of real-world deepfake challenges. However, despite its scale and variability, DeeperForensics focuses on a single manipulation method, face-swapping, which limits its applicability for evaluating detection methods aimed at broader types of manipulations like facial-reenactment or lip-syncing. Table 1 provides a quantitative comparison of XDF with the aforementioned existing deepfake datasets.

2.2 Deepfake Detection Methods

Detecting deepfakes has become a crucial challenge, and several approaches have been proposed, ranging from traditional CNNs to more advanced architectures leveraging temporal and attention mechanisms. Earlier works like Li and Lyu's [12] CNN-based model focused on identifying face-warping artifacts, while Güera and Delp [8] incorporated temporal features by combining CNNs with Long Short-Term Memory (LSTM) networks to analyze sequential frames. Afchar et al. [1] introduced mesoscopic approaches using CNNs and residual neural networks (ResNets) to capture fine-grained tampering evidence in facial images.

With the rise of more sophisticated deepfake techniques, approaches evolved to handle multiple types of manipulation. For instance, Wang and Dantcheva (2019) employed 3D CNNs to detect deepfakes that involve not only face-swapping, but also facial-reenactment and neural texture manipulations. Wodajo and Atnafu [26] combined CNNs with Vision Transformers (ViTs) to capture both spatial and temporal patterns. At the same time, models like Cozzolino et al.'s [4] ID-Reveal and person-of-interest (POI) deepfake detector [3] have pushed boundaries by focusing on personalized detection, examining whether the biometric identity in a video remains consistent over time, rather than just addressing the binary question of whether a video is fake or not.

Recent contributions further illustrate advancements in the field. Zhuang et al. [30] introduced UIA-ViT, an unsupervised ViT-based approach to detect intra-frame inconsistencies without requiring pixel-level annotations. Moreover, El-Gayar et al. [6] utilized a graph neural network-based framework combined with CNNs for deepfake detection, achieving high accuracy on benchmark datasets. Finally, Luan et al. [14] proposed an interpretable deepfake detection model using frequency spatial Transformers, excelling in generalizing to unknown forgery types.

Table 1: Quantitative comparison of XDF with existing deepfake datasets.

Dataset	Real Videos	Fake Videos	Total Videos	Total Subjects	Manipulation Methods	Real Source
UADFV	49	49	98	49	1	YouTube
Deepfake-TIMIT	640	320	960	32	2	VidTIMIT
FF++	1,000	4,000	5,000	1,000	4	YouTube
Celeb-DF	590	5,639	6,229	59	1	YouTube
DFDC	23,654	104,500	128,154	960	8	Self-Recording
DeeperForensics-1.0	10,000	50,000	60,000	100	1	Self-Recording
XDF (Ours)	20,000	60,000	80,000	195	6	YouTube

Table 2: Comparison of datasets with respect to manipulation types and identity verification support.

Dataset	Face-Swapping (Samples)	Facial-Reenactment (Samples)	Lip-Syncing (Samples)	Supports Identity Verification
UADFV	✓(49)	✗	✗	✗
Deepfake-TIMIT	✓(320)	✗	✗	✓
FF++	✓(2,000)	✓(2,000)	✗	✗
Celeb-DF	✓(5,639)	✗	✗	✗
DFDC	✓(128,154)	✗	✗	✓
DeeperForensics-1.0	✓(60,000)	✗	✗	✓
XDF (Ours)	✓(20,000)	✓(20,000)	✓(20,000)	✓

Despite these strides in deepfake detection, a key challenge remains the generalizability of models across diverse manipulation techniques and datasets. Many methods achieve high accuracy on specific datasets but struggle with unseen or more complex manipulations. This highlights the need for datasets that encompass a wide range of manipulation types and identities to enhance model robustness. Unlike prior datasets that focus mainly on face-swapping (Table 2), XDF meets this need by incorporating various manipulation techniques and supporting both binary classification and identity verification tasks, thereby equipping models to better detect a broader spectrum of deepfake techniques and improving their generalizability.

Additionally, our proposed dataset addresses this limitation by incorporating a large number of pristine videos, making it highly suitable for self-supervised and contrastive learning approaches, such as those proposed in [3, 4]. These techniques have shown promising results in improving cross-dataset generalization in recent years, potentially mitigating the performance drop typically observed when a model is trained and tested on different datasets.

3 Methodology

3.1 Pristine Video Collection

To generate a diverse and representative dataset of pristine facial segments, we collected videos from YouTube, featuring 195 different celebrities from various ethnic backgrounds and age groups. This approach ensures that the dataset encompasses a wide range of facial features and expressions, providing a robust foundation for deepfake generation. The detailed steps that outline the process for curating this dataset are presented in Figure 1.

As a first step to retrieve high quality videos featuring an individual “X”, we perform a keyword-based search using the format “X interview” or similar. This method retrieves interview-style videos, which are likely to feature extended appearances of the person of interest. Once a video is downloaded, we extract frames and apply a face detection algorithm to each frame. For this task, we utilize a pretrained version of YOLOv8n [18], specifically finetuned for detecting human faces. The model processes each frame individually, detecting all faces present. To ensure that only frames featuring a single person per frame are retained, we apply a masking process based on the face detection results. Specifically, frames with more than one detected face are discarded. Consecutive frames where a single face is consistently detected are grouped to form continuous segments. This step allows for the creation of video segments that isolate the appearance of the target individual.

The continuous segments formed in the previous step are further segmented based on duration to ensure uniformity across the dataset. Specifically, for each segment, if it exceeds 5 seconds, it is divided into equal-length subsegments, each with the aforementioned duration. Segments that don’t meet this criterion are discarded as they are considered too short for meaningful use in training or analysis.

Finally, the resulting subsegments are subjected to manual filtering to ensure that only high-quality data is included in the final dataset. Each segment is reviewed to ensure that it features the person whose name was originally searched for. Any segments containing individuals other than the target person are discarded. Furthermore, segments that feature static images, where the individual’s face does not exhibit any movement (i.e., a static photograph), are also filtered out. Through this multi-step process, we generate a curated dataset of video segments that reliably contain facial appearances of a specific individual.

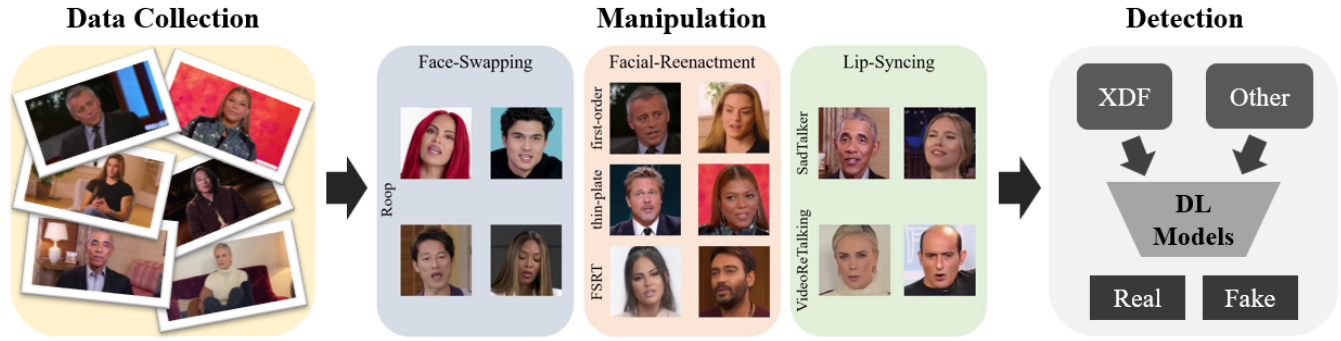


Figure 1: Overview of the proposed methodology.

3.2 Deepfake Video Creation

To generate deepfake videos based on the dataset collected through the previous process, we employ a systematic approach aimed at producing high-quality results. Our deepfake creation pipeline is composed of several key stages as shown in Figure 1. The following outlines the steps in this process.

For each video segment, we begin by tracking the face of the target individual throughout the video. Using the YOLOv8n-face model, we detect the face in each frame and store the bounding box coordinates for the upper left and bottom right corners. These coordinates are maintained in a list, ensuring accurate tracking of the face throughout the video sequence.

Once the face is tracked across all frames, we compute the minimum coordinates of the top-left corners and the maximum coordinates of the bottom-right corners of the bounding boxes throughout the video. This ensures that the bounding box fully captures the face for the entire duration of the segment. To account for minor movements and ensure the entire face is consistently captured, the bounding box is enlarged by a factor of 1.3. This scaling introduces some margin around the face as required by some of the deepfake creation tools used in later stages.

Using the scaled bounding box coordinates, we proceed to video cropping, isolating the facial region from the original frames. However, if any part of the scaled bounding box exceeds the original video frame boundaries, the affected frames are discarded from further processing. This step ensures that only valid regions, fully contained within the frame, are used for subsequent deepfake creation, preventing potential artifacts or distortions caused by incomplete crops.

To create deepfakes, we establish source-target pairs in the following way: In the case of face-swapping, the source refers to the person whose face will be swapped into the target video. In the case of facial-reenactment, the source refers to the person whose facial movements will be transferred to the target video. Finally, in the case of lip-syncing, the source refers to the person whose audio will be used to animate the lips of the target identity. In all cases, the target refers to the person whose identity is being altered or manipulated in the video.

Some of the deepfake generation models used require a single frame from the source (source image) and a video from the target

identity. To enhance the quality of the deepfake, we employ a pose-matching technique. First, we extract five key facial landmarks, such as eyes, nose and mouth corners, from both the source video and the first frame of the target video. The Euclidean distance between corresponding keypoints in the source and target frames is computed to evaluate pose similarity. The frame from the source video that minimizes this distance is selected as the source image. This approach ensures that the relative facial pose of the source closely matches that of the target video, leading to more realistic deepfakes.

With the source-target pairs established and the appropriate source image selected, the final step involves generating the deepfakes. This is accomplished using various deepfake generation tools. These tools typically take the source image and the target video as inputs, applying advanced techniques to produce the final deepfake.

By following this deepfake creation pipeline, we ensure that the generated deepfake videos exhibit high visual fidelity and alignment between the source and target identities. The process incorporates careful tracking, pose matching and rigorous frame selection, contributing to the overall quality of the generated deepfake videos.

3.3 Deepfake Generation Tools

In recent years, numerous tools for creating deepfakes have emerged, leveraging advancements in computer vision and generative models. These tools can be broadly categorized into three main types: face-swapping, where a source person’s face is transferred onto the target person’s head; facial-reenactment, where the expressions and pose of the source person are used to manipulate the target person’s facial movements; and lip-syncing, where an audio track is used to manipulate the mouth movements of the target person to match the source person’s speech. To ensure our dataset is as generalizable as possible, we have included at least one tool from each deepfake generation category. This approach captures a diverse range of manipulation techniques.

3.3.1 Face-Swapping. The tool we used for face-swapping is Roop [22], which implements the face-swapping model provided by InsightFace. Roop produces highly realistic deepfake results while offering a significant advantage over other commonly used face-swapping tools like DeepFaceLab and FaceSwap. Unlike these alternatives, which often require extensive training and separate models for each identity, Roop utilizes a single pre-trained model that

can be applied across multiple identities. This eliminates the time-consuming process of training individual autoencoders for different faces, allowing for the rapid creation of multiple deepfakes in a much shorter time frame.

3.3.2 Facial-Reenactment. We implement three tools for facial-reenactment. All tools require a source image and a driving video that manipulates the expressions and pose of the source image.

The first facial-reenactment tool is the first-order motion model [25]. This method utilizes self-learned keypoints and local affine transformations to capture motion dynamics. Also, it introduces an occlusion-aware generator which estimates occlusion masks to identify and infer object parts not visible in the source image.

The second tool is called thin-plate spline motion model [29]. It uses a keypoint detector to predict keypoints from a driving and a source image, allowing the creation of thin-plate spline transformations. Furthermore, it utilizes an background motion predictor, a dense motion network, and an inpainting network to effectively capture and reconstruct complex motions, and manage possible occlusions.

The third tool is called FSRT [20]. This model generates realistic facial-reenactment results by effectively transferring expressions from a driving image to a target through the following procedure: First, keypoints and facial expressions are extracted to learn a latent representation from the source images. This representation is then used to create the input for the Patch CNN, which is subsequently processed by a Transformer-based architecture. Additionally, the Transformer’s decoder is conditioned on the driving keypoints and expressions, ensuring precise guidance during the generation process.

It is important to note that while there are other tools in the literature that employ facial-reenactment methods and achieve better results than those used in this study, these tools are not publicly available.

3.3.3 Lip-Syncing. For lip-syncing, we employ two models, SadTalker [28] and VideoRetalking [2].

The SadTalker model leverages a 3D Morphable Model (3DMM) to extract and generate coefficients for both pose and facial expressions directly from audio input. These coefficients drive the dynamics of the talking head by providing fine-grained control over facial movements. Furthermore, the model introduces a novel 3D-aware facial rendering pipeline, which ensures the generation of high-quality, synchronized talking heads.

On the contrary, VideoRetalking starts by cropping the facial area and extracting pose and expression coefficients to create the 3DMM. These coefficients are then utilized to generate a video with a consistent expression across all frames. Next, a lip-syncing network, designed with an hourglass architecture and integrated audio modulation, synthesizes realistic lower faces. To further elevate visual quality, an identity-preserving face enhancement network is applied, ensuring high-resolution outputs.

4 Experiments and Results

4.1 Experimental Setup

We evaluate the performance of three DL models commonly used for deepfake detection: MesoNet [1], ResNet18, and EfficientNet-B7, a model that was part of the winning solution of DFDC [23]. Each model is trained on several datasets: the proposed XDF dataset, two established datasets (FF++ and Celeb-DF), and two combined datasets (XDF + FF++ and XDF + Celeb-DF). The combined datasets enable us to evaluate the effectiveness of the proposed XDF dataset when used as an additional training resource and assess its potential for aiding model generalization across different test sets. For FF++ and Celeb-DF, we use the official splits provided by the dataset authors, while for XDF, 75% of the total dataset is used for training, and the remaining 25% is used for testing.

4.1.1 Training Parameters. For MesoNet, the model was trained from scratch using randomly initialized weights. In contrast, for both ResNet18 and EfficientNet-B7, we initialized the models with pre-trained weights from ImageNet.

All models were trained for a total of 10 epochs. The initial learning rate was set to 0.001, and a cosine annealing scheduler was applied to progressively reduce the learning rate during training. The batch size was set to 512 when training MesoNet and ResNet18, and 32 when training EfficientNet-B7 due to memory constraints. We used the Adam optimizer for all experiments, with binary cross-entropy as the loss function, given the binary nature of the deepfake detection task.

4.1.2 Data Preprocessing. To ensure consistency across datasets, the following preprocessing steps are applied prior to training each deepfake detection model. First, facial regions are detected in each frame using a YOLOv8n-based face detector. This model outputs the bounding box of the detected face along with five facial keypoints, specifically the positions of the eyes, nose and mouth corners.

Next, the face is horizontally aligned using the coordinates of the detected eye keypoints. This alignment step, commonly employed in facial analysis tasks, reduces variability in pose and improves the robustness of the detection model. Once aligned, each face is cropped and resized to a resolution of 224×224 pixels. The resized facial images serve as the final input for training.

4.2 Results

In our analysis, we employed accuracy and area under the curve (AUC) as key metrics to evaluate the performance of our dataset in deepfake detection. Accuracy serves as a fundamental measure of a model’s ability to correctly classify instances, providing insights into overall performance. AUC, on the other hand, evaluates the model’s ability to distinguish between classes across various threshold settings.

In Tables 3 and 4, we showcase the performance metrics for our dataset, along with corresponding metrics for FF++ and Celeb-DF. This comprehensive comparison facilitates a clearer understanding of the performance variations across different datasets and model architectures in the context of deepfake detection.

The results reveal several key insights. First, when models are trained and tested on the same dataset, we observe consistently high performance. For example, EfficientNet-B7 achieves an AUC of

Table 3: AUC (%) comparison of each detection model trained and tested on the proposed versus other datasets.

Train \ Test	MesoNet			ResNet18			EfficientNet-B7		
	FF++	Celeb-DF	XDF	FF++	Celeb-DF	XDF	FF++	Celeb-DF	XDF
XDF	57	61	95	60	68	97	62	71	99
FF++	92	69	60	97	70	68	99	71	72
XDF + FF++	90	70	98	98	71	100	99	77	100
Celeb-DF	61	99	62	63	99	63	64	100	53
XDF + Celeb-DF	61	98	94	65	100	100	64	100	99

Table 4: Accuracy (%) comparison of each detection model trained and tested on the proposed versus other datasets.

Train \ Test	MesoNet			ResNet18			EfficientNet-B7		
	FF++	Celeb-DF	XDF	FF++	Celeb-DF	XDF	FF++	Celeb-DF	XDF
XDF	55	73	92	57	72	95	58	71	99
FF++	88	69	54	94	73	52	99	72	50
XDF + FF++	87	77	95	96	78	100	96	79	100
Celeb-DF	41	95	39	37	98	35	40	99	35
XDF + Celeb-DF	46	94	95	38	99	98	44	99	99

99% on FF++ and 100% on Celeb-DF when trained and tested on the same dataset. However, when models trained on one dataset are tested on a different dataset, their performance drops significantly. For instance, ResNet18 trained on FF++ and tested on XDF achieves only a 68% AUC, a substantial decrease from its 97% performance on FF++ alone. This trend highlights a generalization issue common in deepfake detection models, where high intra-dataset performance does not readily translate to cross-dataset scenarios.

A further analysis of XDF’s impact underscores another critical observation: XDF presents a particularly challenging dataset for detection when models are trained on FF++ or Celeb-DF. For instance, EfficientNet-B7, when trained on Celeb-DF, achieves only a 53% AUC on XDF. Similarly, MesoNet trained on FF++ achieves a lower AUC of 60% on XDF, illustrating the increased complexity of detecting deepfakes within the XDF dataset.

Finally, one of the most noteworthy contributions of this study is the performance boost observed when XDF is used as an additional training resource. When models are trained on combined datasets (e.g., XDF + FF++), cross-dataset performance improves substantially. For instance, EfficientNet-B7 trained on XDF + FF++ achieves an AUC of 77% on Celeb-DF, compared to 71% when trained on FF++ alone. This trend is also evident in accuracy results: EfficientNet-B7 trained on XDF + FF++ achieves 79% accuracy on Celeb-DF, outperforming the 72% obtained with FF++ alone.

This improvement in cross-dataset generalization is consistent across other models. For example, MesoNet trained on XDF + FF++ achieves an accuracy of 77% on Celeb-DF, up from 69% when trained on FF++ alone. ResNet18 similarly benefits from the combined training set, achieving 78% accuracy on Celeb-DF with XDF + FF++, compared to 73% with FF++ alone. These findings suggest that incorporating XDF as part of the training set enables models to generalize more effectively to other datasets, underscoring the value of the proposed dataset for improving robustness in deepfake detection.

5 Discussion

Creating a large-scale dataset like the one proposed in this study could contribute significantly to the research community by providing a robust, publicly accessible resource for developing and benchmarking deepfake detection techniques. As generative models become increasingly sophisticated, traditional methods struggle to keep up with the evolving complexity of deepfakes. By curating and sharing such datasets, researchers would have the opportunity to rigorously test and refine their algorithms under realistic and diverse conditions, potentially enabling the development of more accurate and resilient detection systems. Moreover, the scale of the dataset could support the training of more complex models, such as Transformers and graph-based networks, which require substantial data to learn effectively. This type of contribution has the potential to advance the state-of-the-art in deepfake detection while encouraging reproducibility and comparability across relevant studies.

Several key factors must be considered when developing a dataset for deepfake detection to ensure its comprehensiveness and effectiveness. One crucial aspect is the diversity of demographics within the dataset. Deepfakes have the potential to target individuals across different demographic groups, making it essential to include a broad range of characteristics such as gender, ethnicity and age. This demographic diversity ensures that detection models are fair and capable of performing well across various population groups, mitigating biases that could disproportionately affect underrepresented communities.

Another challenge in dataset creation is the inclusion of subjects in extreme poses or where facial occlusion occurs due to glasses, hair or hands. Current deepfake generation techniques often struggle with these conditions, producing poor-quality or unusable results. Incorporating these cases is important to reflect real-world scenarios, where deepfakes might be applied in complex and varied contexts.

This will ultimately enhance the robustness of detection models by exposing them to more diverse and challenging examples.

Moreover, it is essential to include a variety of deepfake creation techniques in the dataset. Deepfake technologies evolve rapidly, and capturing different manipulation methods, such as face-swapping, facial-reenactment and lip-syncing, ensures that the dataset remains relevant against both current and emerging threats. This diversity enables the development of more generalized models that can detect a range of tampering methods.

While some datasets may opt to preserve deepfakes in their original, unaltered form, applying post-processing techniques, such as blurring facial boundaries or performing color correction, could significantly enhance the realism of the generated deepfakes. These techniques help reduce visible artifacts that might make detection easier but unrealistic, thus allowing models to train on data that better represents the types of deepfakes encountered in real-world scenarios.

The creation of such a dataset also requires substantial computational resources. The costs increase with the size and complexity of the dataset, as well as the video resolution and the specific deepfake generation methods used. Researchers must plan for these computational demands, ensuring they have access to the necessary infrastructure to generate and process high-quality deepfakes efficiently.

Beyond dataset creation, an ongoing challenge in the field lies in developing more advanced deepfake detection methods capable of handling increasingly realistic forgeries. The rapid evolution of generative models, such as generative adversarial networks and diffusion models, has resulted in forgeries that are almost indistinguishable from authentic media to both the human eye and basic detection algorithms [24]. Researchers must continue to innovate, exploring novel architectures and techniques, including spatiotemporal analysis, facial landmark tracking and perceptual quality assessment, to stay ahead of forgers. Additionally, there is a need to develop hybrid models that combine different deepfake detection techniques to enhance detection accuracy. Cross-modal methods that analyze audio, video and textual data simultaneously may also prove more effective in detecting inconsistencies and subtle signs of manipulation.

Another open problem is the lack of generalization across deepfake detection models. Often, models perform well on specific datasets but fail to generalize when tested on new or unseen forgeries. Combining multiple datasets during training could help address this issue, enabling models to learn diverse patterns and reducing overfitting to one particular dataset. This approach would enhance models' ability to detect deepfakes in the wild, where forgeries are likely to be more diverse. Researchers should also focus on domain adaptation techniques, transfer learning and data augmentation strategies that improve model robustness and generalizability. These methods will be crucial for developing models capable of real-world deepfake detection, where data is not always clean or uniformly distributed.

Ethical considerations are paramount in deepfake research, particularly around the fair use of subject data and ensuring that the datasets and models are used responsibly. The creation of datasets involving real individuals, whether celebrities or ordinary citizens, raises concerns about consent, privacy and the potential misuse of personal data. It is vital for researchers and institutions to develop stringent guidelines for obtaining and handling subject data ethically.

In addition, there is a growing need for more transparent and interpretable deepfake detection models to foster trust among users and stakeholders. Ensuring that models can provide clear explanations of their decisions will be key in both technical and ethical contexts, particularly in sensitive areas such as media verification, law enforcement and court proceedings.

Finally, public awareness and education also play an essential role in combating the spread of deepfakes. While technological solutions are critical, the general public must be informed about the potential risks of deepfakes, how to recognize them, and the steps they can take to verify media authenticity. Researchers, media organizations and technology companies should collaborate to develop tools and resources that empower individuals to detect and report potential deepfakes. By addressing these limitations, the research community can continue to advance the field. These efforts will not only push the boundaries of what is technologically possible but also contribute to a safer, more informed society.

6 Conclusions

The increasing sophistication of deepfake technologies necessitates the availability of high-quality datasets that can support the development and evaluation of advanced detection models. This study presents a comprehensive dataset of real and manipulated videos, filling a critical gap in the existing resources for deepfake detection. By providing a diverse and extensive collection of video content, this dataset equips researchers with the tools needed to enhance the robustness and accuracy of their detection algorithms. In addition to its practical contributions, this work provides key insights needed for the creation of effective deepfake datasets and identifies several pressing challenges within the field of deepfake detection, offering guidance for future research. By encouraging further inquiry into these areas, the research community can better understand the implications of synthetic media and foster the development of effective solutions. Ultimately, this work aims to contribute to the broader understanding and management of deepfake technologies, paving the way for safer digital environments.

Acknowledgments

This research has been supported by the European Commission funded program EITHOS, under Horizon Europe Grant Agreement 101073928. The computational resources were granted with the support of GRNET.

Data Availability

The dataset is available at: <https://zenodo.org/records/13968987>.

Ethics Statement

The dataset will be made available solely to academic institutions for non-commercial research purposes. The collection and generation of the dataset comply with YouTube's fair use policy, as (1) the material is used in a transformative way for research and educational purposes, (2) only short 5-second segments from the original videos are included, and (3) this use does not interfere with the copyright holder's ability to profit from their original work.

References

- [1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (2018), 1–7.
- [2] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. 2022. VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild. In *Proceedings of the SIGGRAPH Asia 2022 Conference*. 1–9.
- [3] Davide Cozzolino, Matthias Nießner, and Luisa Verdoliva. 2023. Audio-Visual Person-of-Interest DeepFake Detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2023), 943–952.
- [4] Davide Cozzolino, Andreas Rossler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. 2021. ID-Reveal: Identity-aware DeepFake Video Detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 15088–15097.
- [5] Brian Dolhansky, Joanna Bitton, Ben Pfau, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Căntăn Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv: Computer Vision and Pattern Recognition* (2020).
- [6] M. M. El-Gayar, Mohamed Abouhawwash, Sameh S. Askar, and Sara Sweidan. 2024. A novel approach for detecting deep fake videos using graph neural network. *Journal of Big Data* 11 (2024), 1–27.
- [7] Liang Yu Gong and Xue Jun Li. 2024. A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges. *Electronics* (2024). <https://api.semanticscholar.org/CorpusID:267369811>
- [8] David Guera and Edward J. Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (2018), 1–6.
- [9] Liming Jiang, Wayne Wu, Ren Li, Chen Qian, and Chen Change Loy. 2020. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 2886–2895.
- [10] A. Kaur, A. N. Hoshary, V. Saikrishna, S. Firmin, and F. Xia. 2024. Deepfake video detection: challenges and opportunities. *Artif. Intell. Rev.* 57 (2024), 159. doi:10.1007/s10462-024-10810-6
- [11] Pavel Korshunov and Sébastien Marcel. 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. *ArXiv abs/1812.08685* (2018).
- [12] Yuezun Li and Siwei Lyu. 2018. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *CVPR Workshops*.
- [13] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2019. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 3204–3213.
- [14] Tao Luan, Guoqing Liang, and Pengfei Peng. 2024. Interpretable DeepFake Detection Based on Frequency Spatial Transformer. *International Journal of Emerging Technologies and Advanced Applications* (2024).
- [15] Mekhail Mustak, Joni O. Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K. Dwivedi. 2023. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research* (2023).
- [16] Amal Naitali, Mohammed Ridouani, Fatima Salahdine, and Naima Kaabouch. 2023. Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions. *Comput.* 12 (2023), 216.
- [17] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Viet Quoc Pham, and Cu Nguyen. 2022. Deep learning for deepfakes creation and detection: A survey. *Comput. Vis. Image Underst.* 223 (2022), 103525.
- [18] Derron Qi. 2023. YOLOv8-Face: Face Detection Model. <https://github.com/derronqi/yolov8-face>. Accessed: 2024-09-21.
- [19] Md. Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. 2022. Deepfake Detection: A Systematic Literature Review. *IEEE Access* 10 (2022), 25494–25513.
- [20] Andre Rochow, Max Schwarz, and Sven Behnke. 2024. FSRT: Facial Scene Representation Transformer for Face Reenactment from Factorized Appearance, Head-Pose, and Facial Expression Features. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 7716–7726.
- [21] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 1–11.
- [22] s0md3v. 2024. Roop: One-click deepfake face-swapping tool. <https://github.com/s0md3v/roop>. Accessed: 2024-10-10.
- [23] Selim Seferbekov. 2020. DFDC Deepfake Challenge Solution. https://github.com/selimset/dfdc_deepfake_challenge. Accessed: 2024-10-10.
- [24] Jia-Wen Seow, Mei Kuan Lim, Raphaël C.-W. Phan, and Joseph K. Liu. 2022. A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing* 513 (2022), 351–371.
- [25] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and S. Tulyakov. 2021. Motion Representations for Articulated Animation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 13648–13657.
- [26] Deressa Wodajo and Solomon Atnafu. 2021. Deepfake Video Detection Using Convolutional Vision Transformer. *ArXiv abs/2102.11126* (2021).
- [27] Xin Yang, Yuezun Li, and Siwei Lyu. 2018. Exposing Deep Fakes Using Inconsistent Head Poses. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), 8261–8265.
- [28] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xiaodong Shen, Yu Guo, Ying Shan, and Fei Wang. 2022. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 8652–8661.
- [29] Jian Zhao and Hui Zhang. 2022. Thin-Plate Spline Motion Model for Image Animation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 3647–3656.
- [30] Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. 2022. *UIA-ViT: Unsupervised Inconsistency-Aware Method based on Vision Transformer for Face Forgery Detection*. Springer, 391–407.