2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

# INDIFACE: Illuminating India's Deepfake Landscape with a Comprehensive Synthetic Dataset

Kartik Kuckreja[1], Ximi Hoque[2], Nishit Poddar[1], Shukesh Reddy[1], Abhinav Dhall[2] and Abhijit Das[1]
[1] Birla Institute of Technology and Science, Pilani, Hyderabad Campus, India
[2] Indian Institute of Technology, Ropar, India
abhijit.das@hyderabad.bits-pilani.ac.in

***Abstract*— Due to the recent progress in Deepfake generation, several datasets and manipulation techniques have been proposed in the recent literature with various effective face-swap and face-reenactment methods. Deepfake is an emerging threat to society and government as it can jeopardize law enforcement and cause personal loss. Investigations in the literature established that demographic variation had impacted the performance of Deepfake detection. To date, Deepfake detection has not been studied in the Indian context; hence, in this work, we proposed a Deepfake dataset INDIFACE entirely with Indian subjects. We have collected 101 original videos and used two different manipulation techniques for Deepfake generation. We provide detailed benchmarking with state-of-the-art methods on Deepfake datasets, showcasing that the existing model is insufficient to detect Deepfake detection for the Indian scenario. Hence, more attention is required to this area of research. The proposed dataset INDIFACE is publicly available at .**

## I. Introduction

Deepfakes are computer-generated videos, images, or audio recordings manipulated using artificial intelligence and machine learning algorithms to create realistic content of a person doing or saying something they did not do. Deepfakes use deep learning techniques like neural networks to manipulate existing content and create something new. The term "Deepfake" is a combination of "deep learning" and "fake". Deepfakes are threats to society because they can be used to spread misinformation, manipulate public opinion, harass individuals, and even blackmail people. They can be particularly dangerous when used in political contexts, where they can be used to damage reputations or influence election outcomes. Deepfakes can be generated in several ways; manipulation methods involve generative adversarial networks (GANs) to create realistic videos. Another method involves using facial detection technology to map a person's face onto a body or to superimpose their face onto an existing video. One common approach is to train an auto-encoder to reconstruct a specific person's face on any given image of a body.

Due to the recent fuel in media tampering techniques such as Deepfake, corresponding manipulation-detection approaches have been developed. Manipulation detection techniques included image-based [24], video-based [11], [25], or jointly audio and video-based [9] approaches. In the context of image-based Deepfake detection to explore the angle of generalizability, second-order local anomaly detection has been used [16] and self-consistency is explored in the work of [27]. Singular frame-based detection techniques ensemble predictions across video frames [31], [7]. While being computationally efficient, they do not exploit the presence of temporal inconsistencies [21]. Hence, video-based generalized deepfake detection has recently gained significance. This is for the detection of temporal inconsistencies concerning the lip movement [1], jitters between frames [18], and optical flow [3]. Identity consistency has also been explored, with [15] using a transformer to identify inconsistency between the inner and outer face region based on a database of known identities. Authors of [2] model behaviors of world leaders from recorded stock footage and identify behavioral inconsistencies in Deepfakes. These techniques usually require prior identity or behavior information about the victim, so they are suited for celebrities but do not scale to civilian victims.

In [5], self-supervised learning (SSL) has been explored by learning adversarial examples for generalized deep fake detection. Further, in the same research direction, a multimodal approach has been adopted in [9], [32] using audio-video analysis. Tolosana et al. [28] reviewed the first and second eras of manipulation techniques, such as DeepFake in 2020 w.r.t. facial regions, and fake detection performance and concluded that the generalization of such detection methods is challenging. In other words, when detection methods, such as those presented, are confronted with adversarial attacks outside of the training set, such networks dramatically drop performance. The challenge of generalization of deep fake is studied in [11], [25].

Considering face analysis, another angle that comes as a challenge for generalizing face analysis is the bias owing to race, age, and gender [10]. This will also be an obvious additional problem for deepfake. To date, the datasets that are developed on Deepfake, such as FaceForensics++ [24] and DFDC [14], do not primarily contain Indian subjects. The DFDC [14] dataset, which has a noteworthy lack of East Asian and Southeast Asian subjects (the proportion of East Asians in the database is 9%, while that of Southeast Asians is 3%), is balanced by the Korean subjects (and the eight Southeast Asians) in KoDF [20]. Building more generalized detection models for practical applications will need combining the complementing racial makeup of different datasets based on different demographics. In this line of research, the KoDF dataset [20] is developed to foster deepfake for Korean

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

TABLE I

QUANTITATIVE COMPARISON OF INDIFACE WITH THE EXISTING PUBLIC DEEPFAKE DETECTION DATASETS.

| Dataset | Real videos | Fake videos | Total Videos | Total Subjects | Sources | Demography Considered |
|---|---|---|---|---|---|---|
| UADFV[30] | 49 | 49 | 98 | 49 | Youtube | × |
| DeepfakeTIMIT[19] | 640 | 320 | 960 | 32 | VidTIMIT | × |
| FF++[24] | 1,000 | 4,000 | 5,000 | N/A | Youtube | × |
| Celeb-DF | 590 | 5639 | 6229 | 59 | Youtube | × |
| DeePhy[22] | 100 | 5040 | 5140 | 100 | Youtube | × |
| DFDC[14] | 23,654 | 104,500 | 128,154 | 960 | Self Recording | × |
| KoDF[20] | 62,166 | 175,776 | 237,942 | 403 | Self-Recording | ✓(Korean) |
| DF-Platter[23] | 764 | 132,496 | 133,260 | 454 | Youtube | × |
| **INDIFACE** | **404** | **1668** | **2072** | **58** | **Youtube and Self captured images** | **✓(Indian)** |

demographics. Recent work in [23] introduced some Indian subjects in their dataset, but an analysis of the performance for the Indian subset was not carried out. To resolve this gap in the literature, we attempt to study the Indian scenario. This study will try to answer the following research questions:

- Is the state-of-the-art techniques on deepfake detection appropriate to detect manipulation of videos containing Indian subjects?
- If not it is not appropriate, are deepfakes different for the Indian scenario?
- How can we mitigate the gap and bring generalization?

Hence, it will be interesting to analyze the performance of state-of-the-art techniques in deepfake detection on Indian subjects. To answer the above questions, we developed a new deepfake dataset entirely using Indian subjects. We experimented with state-of-the-art techniques on deepfake detection to find out if they are ornamental for detecting manipulation in videos for Indian scenarios. However, the next question is which type of task detection is pertinent to this problem.

To this end, our experimental analysis shows that the existing state-of-the-art model could not perform well; further fine-tuning the model has led to respectable detection results.

## II. RELATED WORK

Now, we enlist the publicly available datasets and state-of-the-art detection techniques on Deepfakes.

### A. Deepfake Datasets

The following are the most popular publicly available datasets on Deepfake. Table I enlists the quantitative comparison of the dataset concerning the proposed dataset.

- **FaceForensics++**: The FaceForensics++(FF++) dataset [24] is a large-scale benchmark dataset for face manipulation detection, which was created to help develop automated tools that can detect deepfakes and other forms of facial manipulation. The dataset consists of over 1,000 high-quality videos with over 500,000 frames, generated using various manipulation techniques such as facial reenactment, face swapping, and deepfake generation. The videos in the dataset are divided into four categories, each corresponding to a different manipulation technique: Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Deepfakes use machine learning algorithms to generate realistic-looking fake videos, while Face2Face and FaceSwap involve manipulating a person's facial expressions and identity in a video. NeuralTextures uses a different approach by altering the texture of a face to make it appear different. The dataset includes both real and manipulated videos, with each manipulation technique applied to multiple individuals.
- **DFDC [14]**: The Partnership on AI's Media Integrity Steering Committee, among other well-known companies, Amazon Web Services, Facebook, Microsoft, and others, worked with academics to create the DeepFake Detection Challenge (DFDC [14]) in 2020. The DFDC [14] project is a substantial endeavor comprising a competition, a sizable dataset, and related academic articles. This dataset, which includes more than 960 people and more than 120,000 movies, is the third largest public Deepfake dataset after KoDF [20] and DF-Platter [23]. The raw clips in DFDC [14] were obtained from several environmental situations, recording a broad range of lighting, audio, and angle circumstances to ensure the diversity and complexity of the collection. Additionally, eight distinct synthesis techniques were used to create the Deepfake films. The DFDC [14] dataset does have certain restrictions, though. Extreme changes in lighting, audio quality, and camera angles were produced through the unguided recording procedure, where individuals filmed themselves. The dataset also shows inconsistent formatting, with individual video clips differing in resolutions and lengths. Additionally, the demographic distribution of participants is uncontrolled. It is important to note that the DFDC [14] dataset might not specifically address the issues unique to Indian demography, given our focus on Deepfake datasets relating to Indians. In our research, we, therefore, intend to investigate and address the special features and difficulties relating to Deepfake in the Indian scenario.
- **KoDF** [20]: The Korean Deepfake Detection Dataset(KoDF [20]) is the largest publicly available

dataset on synthesized videos, encompassing 62,166 unique 90-second-long real clips (62.8 days) and 175,776 different Deepfake clips of at least 15 seconds(30.5 days). Six unique synthesis models are used to create the Deepfake samples. Most participants in KoDF [20] are Koreans to balance the Asian demographics that are underrepresented in the current Deepfake detection databases. Finally, the dataset employs several strategies to control better the distribution of the participant's age, sex, and content data.

- **DF-Platter:** DF-Platter [23] is a large-scale Deepfake dataset consisting of 133,260 videos, each lasting approximately 20 seconds, with an estimated duration of 30.67 days. It is the second-largest dataset in terms of video count, following KoDF [20]. The dataset includes both high-resolution (HR) and low-resolution (LR) Deepfake videos, organized into three sets: Set A, Set B and Set C. Set A contains single-subject Deepfakes, while Sets B and C consist of multi-face Deepfakes involving manipulations of multiple subjects' faces in different ways. The videos in the dataset were gathered in the wild, specifically from YouTube, and vary in gender, orientation, skin tone, face size, lighting conditions, background, and whether or not occlusion is present. When hands, hair, eyewear, or any other object obscures a portion of the source or target face, it is said to be occluded. The dataset's diversity is achieved by applying three distinct techniques - FSGAN, FaceShifter, and FaceSwap - for video generation. The DF-Platter [23] dataset is balanced across resolution and gender, in contrast to the majority of publicly available datasets that are unbalanced across various attributes like age, gender, and skin tone.

We can conclude from the literature that demographics have a huge influence on deep fake detection, and in this context, the Indian demographic is not explored much except for DF-Platter [23] where Indian subjects were considered, though separate analyses on the Indian subjects or any cross-demographic and cross-manipulation experiments were not carried out. To investigate the gap, we attempted to forge this work.

### B. Deepfake Detection

Due to the recent progress in media tampering techniques such as Deepfake, corresponding *manipulation-detection approaches* have been developed. Manipulation detection techniques included image-based [24], video-based [11], [25], [13], [4], or jointly audio and video-based [9] approaches. In the context of image-based Deepfake detection to explore the angle of generalizability, second-order local anomaly detection has been used [16] and self-consistency is explored in the work of [27]. Singular frame-based detection techniques ensemble predictions across video frames [31], [7]. While computationally efficient, they do not exploit the presence of temporal inconsistencies [21].

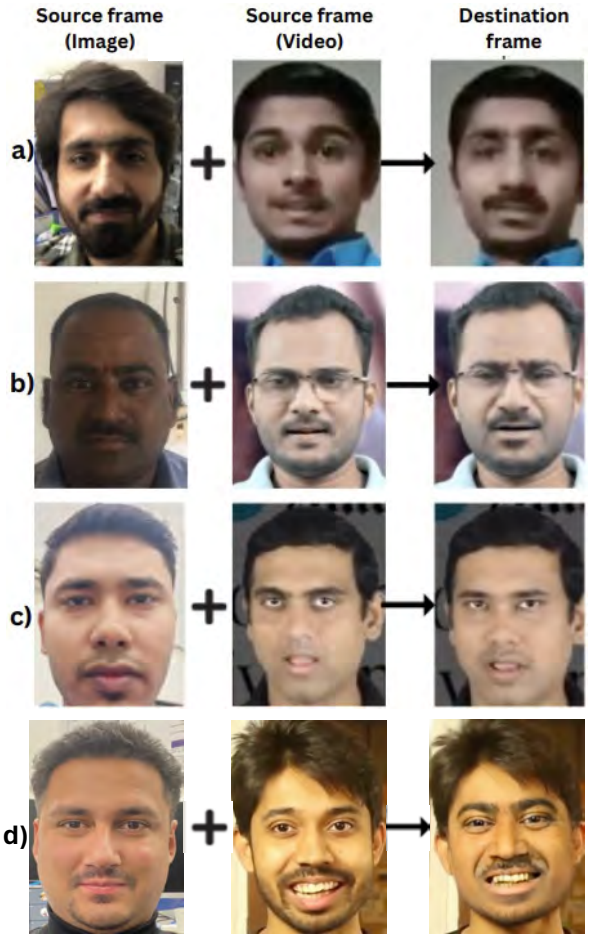


Fig. 1. Deepfakes made using SimSwap [6] (a, b) and Ghost [17] (c,d).

Hence, recently, video-based generalized Deepfake detection has been gaining significance. This is for the detection of temporal inconsistencies concerning the lip movement [1], jitters between frames [18], and optical flow [3]. Identity consistency has also been explored, with [15] using a transformer to identify inconsistency between the inner and outer face region based on a database of known identities. Authors of [2] modeled behaviors of world leaders from recorded stock footage and identified behavioral inconsistencies in Deepfakes. These techniques usually require prior identity or behavior information about the victim, so they are suited for celebrities but do not scale to civilian victims.

## III. Proposed Dataset

In this section, we explain the proposed dataset along with the data collection process, manipulation technique, and process employed to generate fake videos.

### A. Real Clips Collection

*1) Data Collection:* Our data collection process focused on obtaining diverse videos featuring individuals of Indian origin (see Fig. 1 and 2). This section outlines the steps in gathering and preparing the real videos for analysis. We sourced 62 videos exclusively featuring individuals of Indian

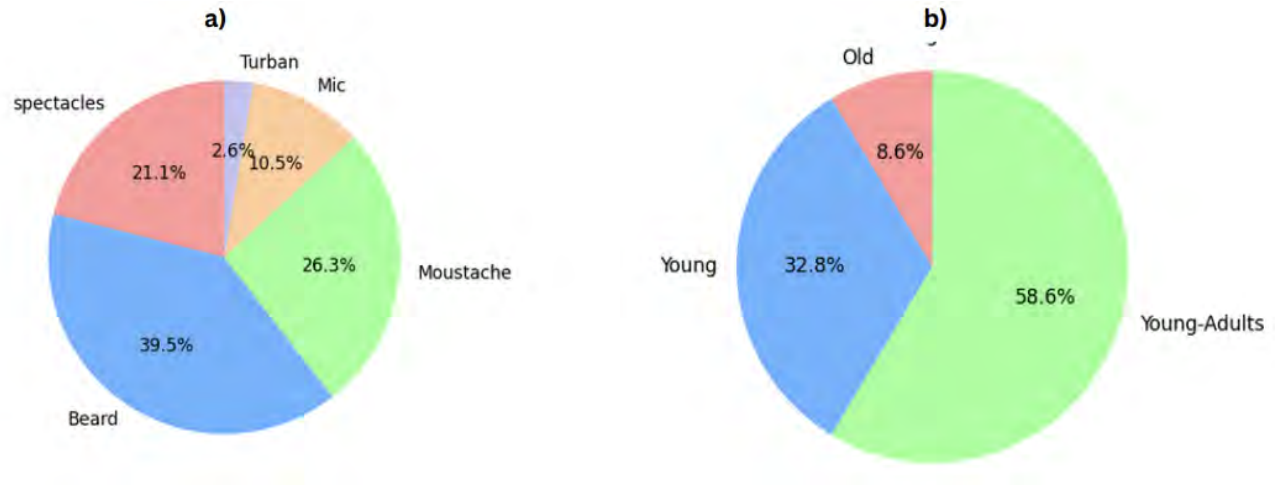


Fig. 2. Characteristics of INDIFACE (a) Different types of occlusions present in the videos. (b) The age spread of the dataset.

origin from YouTube. These videos were chosen to represent various contexts and scenarios, and to better represent the wide-scale presence of Indian demography in terms of skin tone and facial features, resulting in a diverse dataset with content ranging from 2-minute to 1.5-hour clips.

*2) Clip Selection and Preprocessing:* From each of the 62 selected videos, multiple sequences were extracted where the subject's face was visible front-facing for a continuous time frame of at least 10 seconds. These segments were chosen to ensure the dataset's focus on front-facing facial data.

We further processed the selected segments to augment the dataset and mimic real-life scenarios. Each 10-second segment underwent three separate preprocessing techniques:

- **Gaussian Blur**: A Gaussian blur filter was applied to mimic the effect of a slightly out-of-focus camera, introducing visual artifacts often seen in real-world videos.
- **Salt and Pepper Noise**: Random "salt" (white) and "pepper" (black) pixels were added to the clips to simulate visual noise commonly encountered in practical video recordings.
- **Random Shifting Brightness**: The brightness levels of the clips were randomly adjusted to represent different lighting conditions, reflecting the challenges faced by real-world Deepfake detection models.

These pre-processing steps resulted in *404 real videos* in total, comprising 101 original 10-second clips and their respective variations.

### B. Fake Clips Generation

*1) Face Collection:* To represent the diverse population of India, we carefully selected 19 individuals of Indian origin from different regions with varying skin tones and facial structures. The intention was to include a wide range of ethnic and cultural backgrounds to create a comprehensive and inclusive dataset.

*2) Deepfake Video Generation:* To generate the fake clips, we used two manipulation methods, SimSwap [6] and Ghost [17] -one-shot Transfer. Since we manually screened the generated videos, the number of videos from both methods is not the same. Following are the details of the manipulation techniques used.

- **SimSwap [6]** SimSwap [6](Simple Swap) is an efficient face swapping method for generalized and high fidelity face swaps. It uses the ID Injection Module (IIM) to transfer identity information from a source face to a target face, allowing for seamless transfer of identities without underlying differences. It also introduces the Weak Feature Matching Loss, a novel approach to attribute preservation that efficiently retains crucial facial attributes during the swapping process. This loss function implicitly maintains essential characteristics, such as facial expressions, resulting in more realistic and visually appealing results. SimSwap [6]'s superiority over existing methods is demonstrated through extensive experiments on diverse real-world faces, achieving competitive performance in comparison to state-of-the-art methods. To create the synthesized clips, a source image of a person and a driving video is passed to the SimSwap [6] model, which is pre-trained on the VGG-Face2-HQ.
- **Ghost[17]** The Ghost [17] -One Shot Transfer method is an efficient and innovative framework for high-quality face swapping, enabling generalized identity transfer and attribute preservation. It consists of four main components: an identity encoder to extract source identity information, an attribute encoder based on U-Net architecture to extract target attribute features, an AAD generator to combine identity and attribute vectors, and a multiscale discriminator for image quality comparison. A customized loss function is employed to enhance the model's performance, including reconstruction loss, attribute loss, identity loss, and adversarial loss. Additionally, the model incorporates an eye loss to maintain consistent gaze direction in video swaps. It addresses shape mismatches through landmark tracking and en-

sures stability via bounding box smoothing. Extensive experiments demonstrate its effectiveness in producing visually appealing and convincing face swaps. To create our Deepfake clips, we used a Ghost [17] model pre-trained on the VGG-Face2 dataset.

### C. Informed Consent and License Compliance

Before capturing any facial data, we obtained informed consent from each participant, providing clear explanations about the purpose of data collection and how it would be utilized in deepfake video generation. Participants were assured of the confidentiality of their data and that it would be used exclusively for non-commercial research purposes. Only those videos with open Creative Commons licenses were selected for YouTube videos.

## IV. Benchmarking

To benchmark our dataset, we use two models, Selim implementation of EfficientNet [26] and Cross Efficient ViT [8]. We discuss them as follows:

- **Selim (DFDC [14] Winner)** Selim is an implementation of EfficientNet-B7 [26], which won the DFDC [14] challenge on Kaggle. The SOTA method for identifying Deepfake clips largely uses a frame-by-frame classification methodology. It uses MTCNN [29] face detector instead of S3FD because of the quick and memory-efficient performance. The careful configuration of input sizes, customized to video resolutions for optimal processing, is a subtle aspect. The model's generalization abilities are improved via augmentation methods, including isotropic scaling and dropout-based alterations. Adding a margin during preprocessing strengthens the model's resistance to various input sizes. All of these pre-processing steps differentiate it from Cross Efficient ViT [8], and both models currently perform the best on DFDC [14].
- **Combining EfficientNet and Vision Transformers for Video Deepfake Detection** This paper combines the capability of EfficientNet [26] and Vision Transformers. They explore the detection of manipulated faces in videos using a combination of two computational techniques: convolutional and transformer approaches. They also use MTCNN [29], followed by developing two distinct methods: Efficient ViT and Convolutional Cross ViT. These methodologies aim to ascertain the authenticity of a face within the videos. Efficient ViT employs a blend of techniques, while Convolutional Cross ViT examines both small and large facial components to make determinations. Testing across various videos reveals that Convolutional Cross ViT excels in detecting manipulated faces, even when they significantly differ from the original. Cross Convolution ViT, trained on DFDC [14] and FF++ [24], achieves the highest AUC score on DFDC [14] test set.
- **SSAT** [12]: Optimizing ViTs for the primary task and a Self-Supervised Auxiliary Task (SSAT) is surprisingly beneficial for low data scenarios as ours. It explores the appropriate SSL tasks that can be optimized alongside the primary task, the training schemes for these tasks, and the data scale at which they can be most effective. SSAT is a powerful technique that enables ViTs to leverage the unique characteristics of both the self-supervised and primary tasks, achieving better performance than typical ViTs pre-training with SSL and fine-tuning sequentially.

## V. Experiments and Results

In this section, we proceed to discuss the experimental protocol and the different findings that we can draw from this study.

### A. Experimental protocol

The ultimate objective of a Deepfake detection dataset would be to aid in creating a general detection model that excels against various Deepfake situations encountered in the real world. Most studies on Deepfake detection are set up to evaluate the effectiveness of the suggested detection models using a specific Deepfake detection dataset. The target Deepfake detection dataset is assumed to be a good approximation of the distribution of actual Deepfake instances.

For all these experiments, we split INDIFACE into three parts in the ratio of 7:1:2 for training, validation, and testing, respectively. The dataset was parted while keeping all the splits subjects independent, meaning that all the splits contain different source videos with different subjects.

We use MTCNN [29] to first extract the faces from INDIFACE and then do equidistant sampling to get 32 frames per video as our input. We fine-tuned the models for 40 epochs and picked the weights with the least validation loss for testing.

The upcoming evaluation strategy is forged to examine if the available deep fake detection datasets guarantee a suitable level of generality for the Indian subjects and how well they perform when combined and evaluated with data from other domains. Hence, we adopted the following strategy:

- **Zero-shot evaluation:** We DFDC [14] winning model Selim and Cross Efficient ViT [8] model trained on DFDC [14], and FF++ [24] and tested on INDIFACE.
- **Finetuned on INDIFACE:** In this version we finetuned DFDC [14] winning model Selim, SSAT [12] and Cross Efficient ViT [8] model trained on DFDC [14], and FF++ [24] with INDIFACE training anad validation split. Then, the performance of the was evaluated on the test split of INDIFACE.

### B. Results on the benchmarking

The initial zero shot evaluation gave surprisingly low results for SSAT [12], Selim (DFDC [14] Winner) pre-trained on DFDC [14], and Cross Efficient ViT [8] model pre-trained on DFDC [14] and FF++ [24], thus showing that both of these models are not enough to generalize to Deepfakes for Indian scenario (see Table II). However, the performance of the model, when finetuned on INDIFACE, is noteworthy refer

TABLE II

ZERO SHOT EVALUATION OF OUR TEST SET ON PRE-TRAINED CROSS EFFICIENT VIT [8], SSAT [12] AND SELIM(DFDC [14] WINNER)

| Zero Shot | Cross Efficient ViT [8]-Efficient Net | | | Selim(Efficient Net) | | | SSAT | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | AUC score | F1 score | Accuracy | AUC score | F1 score | Accuracy | AUC score | F1 score |
| INDIFACE | 0.56 | 0.74 | 0.60 | 0.69 | 0.79 | 0.72 | 0.20 | 0.50 | 0.06 |
| SimSwap [6] | 0.57 | 0.76 | 0.61 | 0.74 | 0.82 | 0.76 | 0.25 | 0.51 | 0.07 |
| Ghost [17] | 0.69 | 0.84 | 0.72 | 0.69 | 0.92 | 0.72 | 0.47 | 0.55 | 0.04 |

TABLE III

PERFORMANCE OF CROSS EFFICIENT VIT [8], SSAT [12] AND SELIM(DFDC [14] WINNER), PRETRAINED ON DFDC AND FF++ [24], FINETUNED ON INDIFACE.

| Fine-tuned on Base | Cross Efficient ViT [8] | | | Selim(Efficient Net) | | | SSAT | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | AUC score | F1 score | Accuracy | AUC score | F1 score | Accuracy | AUC score | F1 score |
| **INDIFACE** | 0.96 | 0.99 | 0.96 | 1.00 | 0.99 | 1.00 | 0.87 | 0.93 | 0.93 |
| SimSwap [6] | 0.98 | 1.00 | 0.98 | 0.99 | 1.00 | 0.99 | 0.87 | 0.93 | 0.91 |
| Ghost [17] | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.86 | 0.96 | 0.88 |

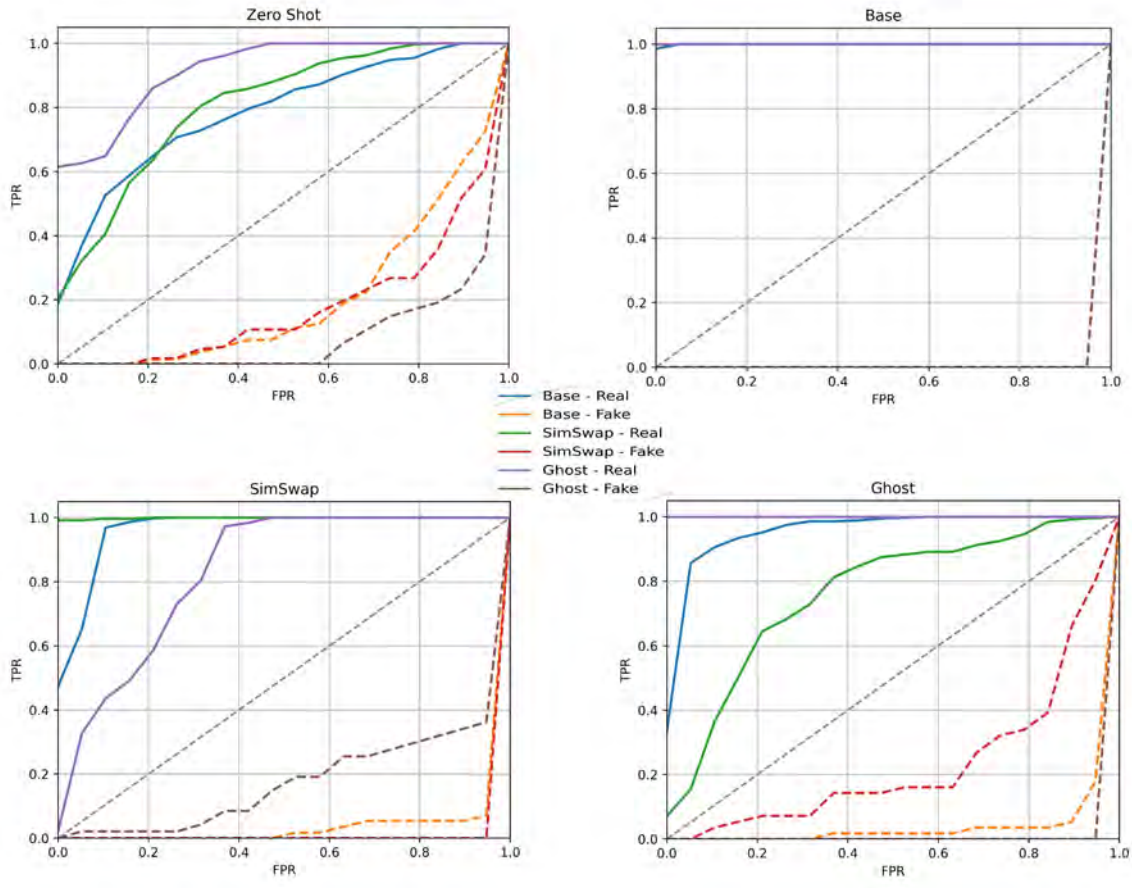


Fig. 3. ROC curves for Selim (DFDC[14] Winner).

TABLE IV

PERFORMANCE OF CROSS EFFICIENT VIT [8], SSAT [12] AND SELIM(DFDC [14] WINNER), PRETRAINED ON DFDC AND FF++ [24], FINETUNED ON SAMPLES FROM SIMSWAP [6].

| Fine-tuned on SimSwap[6] | Cross Efficient ViT [8] | | | Selim(Efficient Net) | | | SSAT | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | AUC score | F1 score | Accuracy | AUC score | F1 score | Accuracy | AUC score | F1 score |
| INDIFACE | 0.50 | 0.91 | 0.56 | 0.69 | 0.97 | 0.74 | 0.69 | 0.86 | 0.77 |
| **SimSwap [6]** | 0.91 | 0.99 | 0.91 | 0.97 | 1.00 | 0.97 | 0.85 | 0.93 | 0.90 |
| Ghost [17] | 0.47 | 0.83 | 0.48 | 0.68 | 0.83 | 0.71 | 0.51 | 0.70 | 0.24 |

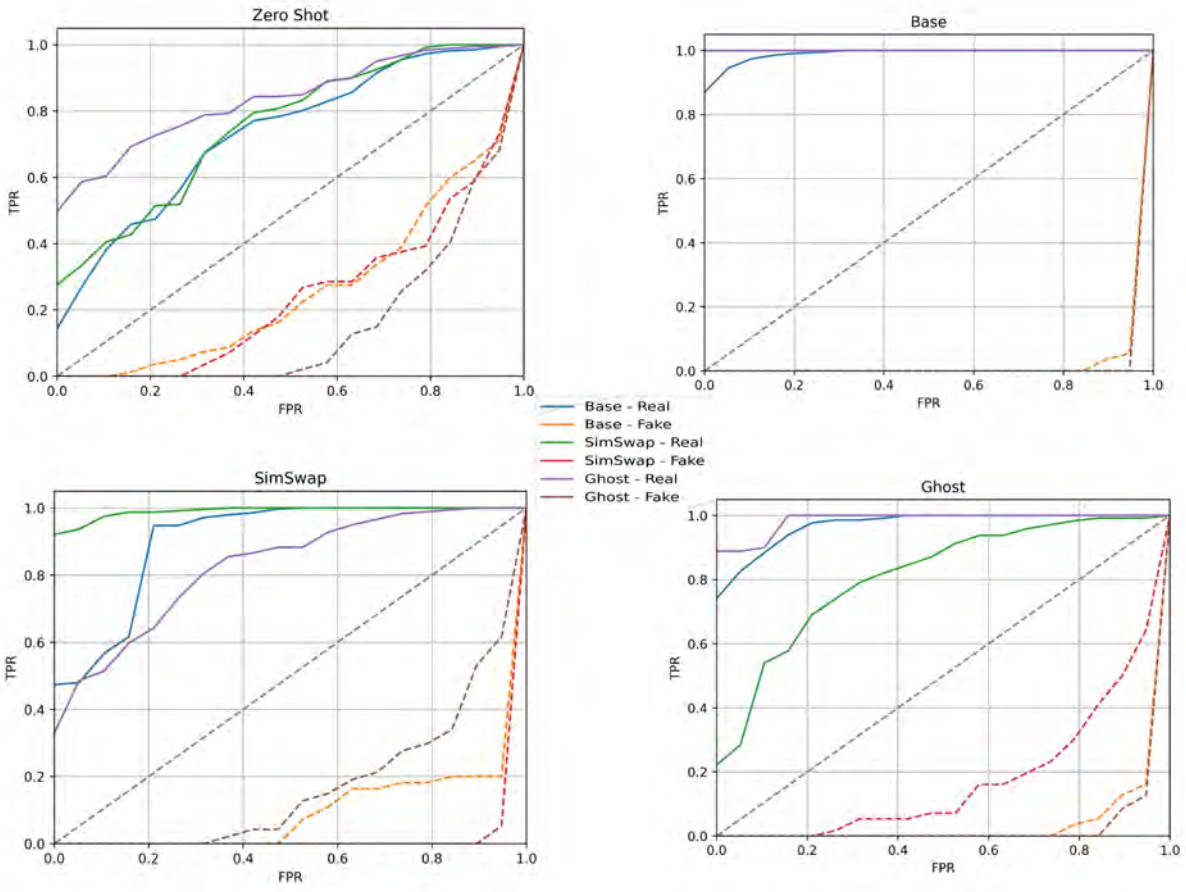


Fig. 4. ROC curves for Cross Efficient ViT [8] efficient net

TABLE V
PERFORMANCE OF CROSS EFFICIENT VIT [8], SSAT [12] AND SELIM(DFDC [14] WINNER), PRETRAINED ON DFDC AND FF++ [24], FINETUNED ON SAMPLES FROM GHOST [17] .

| Fine tuned on Ghost [17] | Cross Efficient ViT [8] | | | Selim(Efficient Net) | | | SSAT | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | AUC score | F1 score | Accuracy | AUC score | F1 score | Accuracy | AUC score | F1 score |
| INDIFACE | 0.64 | 0.97 | 0.70 | 0.44 | 0.96 | 0.49 | 0.39 | 0.69 | 0.43 |
| SimSwap [6] | 0.61 | 0.82 | 0.65 | 0.37 | 0.77 | 0.38 | 0.25 | 0.59 | 0.11 |
| **Ghost [17]** | 0.93 | 0.99 | 0.92 | 0.96 | 1.00 | 0.96 | 0.93 | 0.96 | 0.94 |

to Table III; the same conclusion can be drawn from ROC curves from Fig. 3. 4, and 5.

Further, to study the effect of different methods used to create the Deepfakes, we cross-evaluated the models by finetuning on videos made using a single method and testing the others. The results in Tables IV and V show that the models cannot generalize to deepfakes developed by different methods and are vulnerable to overfitting. The main reason behind this overfitting is that the detection models try to learn from the artifacts present in the frames, which varies according to the methods used. The point is that the models become significantly more resistant to various types of non-domain data when trained on combinations of different methods.

From the experiment results, we can conclude that an ideal Deepfake detection dataset should include a large variety of Deepfake generation methods. We should include a variety of real videos. No existing dataset can generalize to all demographics; combining these datasets should be the way to move forward.

Observing the experimental results, now proceed to answer the question raised in the introduction:

- **Are the state-of-the-art techniques on deep fake detection appropriate to detect manipulation of videos containing Indian subjects?**
  **Answer:** The results show the incapability of the models to generalize to Deepfakes developed for Indian scenarios using different manipulation methods and their vulnerability to overfit. This overfitting is because the detection models try to learn from the artifacts,

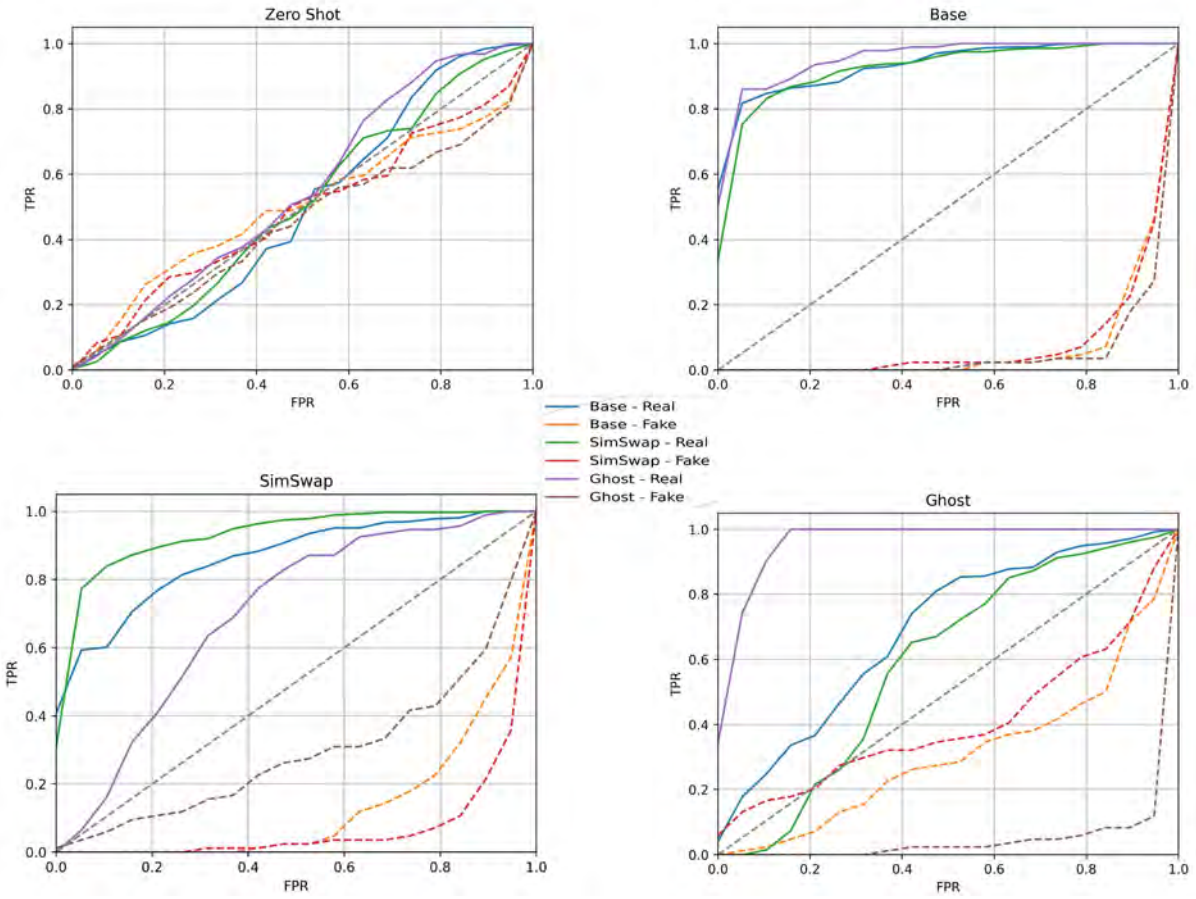


Fig. 5. ROC curves for SSAT

which, in turn, depends on the methods used to make the clips and other covariates.

- **If not it is not appropriate, are deep fakes different for the Indian scenario?**
  **Answer:** The detection models failed because they try to learn from the artifacts, which, in turn, depends on the methods used to make the clips and other covariates. These factors have a big role in maintaining spatiotemporal synergy while generating deep fake.
- **How can we mitigate the gap and bring generalization?**
  **Answer:** Finetuning the model with the new data has given a solution. Hence, some type of meta-learning or incremental learning will be the solution that should be investigated.

## VI. Conclusions

In the recent literature on Deepfake, several datasets and manipulation techniques have been proposed, but they have not considered Indian subjects. This is the first work on Deepfake where only Indian subjects are considered, and the impact on the performance of Deepfake with state-of-the-art Deepfake detection techniques.

We proposed the INDIFACE dataset, which is entirely based on Indian subjects. It consists of 404 real videos and 1668 fake videos generated using two different manipulation techniques. We employed a couple of state-of-the-art deepfake detection techniques for benchmarking. From the experiment conducted it can be concluded that the existing model cannot detect Deepfake detection on Indian subjects until and unless they are finetuned. Hence, it proves from this initial study that research attention is required to achieve a satisfactory performance on Deepfake for Indian subjects.

In future research, we will study more manipulation techniques, enlarge the dataset, and come up with an efficient detection technique for this area of research.

## Acknowledgements

This work was funded by the Institute of Data Engineering, Analytics, and Science (IDEAS) Technology Innovation Hub (TiH) Indian Statistical Institute, Kolkata, under the aegis of National Mission on Interdisciplinary Cyber-Physical Systems (NM-ICPS), Department of Science and Technology (DST), Government of India under the project titled ”Generalized Tampering Detection in Media (GTDM)” and project number OO/ISI/IDEAS-TIH/2023-24/86.

## REFERENCES


[1] S. Agarwal, H. Farid, O. Fried, and M. Agrawala. Detecting deepfake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661, 2020.

[2] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, page 38, 2019.

[3] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[4] P. Balaji, A. Das, S. Das, and A. Dantcheva. Attending generalizability in course of deep fake detection by exploring multi-task learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 475–484, 2023.

[5] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang. 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*, pages 18710–18719, 2022.

[6] R. Chen, X. Chen, B. Ni, and Y. Ge. Simswap: An efficient framework for high fidelity face swapping. In *MM '20: The 28th ACM International Conference on Multimedia*, 2020.

[7] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pages 219–229. Springer, 2022.

[8] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi. Combining efficientnet and vision transformers for video deepfake detection. In S. Sclaroff, C. Distante, M. Leo, G. M. Farinella, and F. Tombari, editors, *Image Analysis and Processing – ICIAP 2022*, pages 219–229, Cham, 2022. Springer International Publishing.

[9] E. Conti, D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, A. Sarti, M. C. Stamm, and S. Tubaro. Deepfake speech detection through emotion recognition: a semantic approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8962–8966. IEEE, 2022.

[10] A. Das, A. Dantcheva, and F. Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[11] A. Das, S. Das, and A. Dantcheva. Demystifying attention mechanisms for deepfake detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7. IEEE, 2021.

[12] S. Das, T. Jain, D. Reilly, P. Balaji, S. Karmakar, S. Marjit, X. Li, A. Das, and M. Ryoo. Limited data, unlimited potential: A study on vits augmented by masked autoencoders. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[13] S. Das, T. Jain, D. Reilly, S. Karmakar, S. Marjit, X. Li, and M. Ryoo. From few to more: Enhancing vit performance on limited data. 2023.

[14] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[15] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9468–9478, 2022.

[16] J. Fei, Y. Dai, P. Yu, T. Shen, Z. Xia, and J. Weng. Learning second order local anomaly for general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20270–20280, 2022.

[17] A. Groshev, A. Maltseva, D. Chesakov, A. Kuznetsov, and D. Dimitrov. Ghost—a new face swap approach for image and video domains. *IEEE Access*, 10:83452–83462, 2022.

[18] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[19] P. Korshunov and S. Marcel. Deepfakes: a new threat to face recognition? assessment and detection, 2018.

[20] P. Kwon, J. You, G. Nam, S. Park, and G. Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10744–10753, 2021.

[21] Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.

[22] K. Narayan, H. Agarwal, K. Thakral, S. Mittal, M. Vatsa, and R. Singh. Deephy: On deepfake phylogeny. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2022.

[23] K. Narayan, H. Agarwal, K. Thakral, S. Mittal, M. Vatsa, and R. Singh. Df-platter: Multi-face heterogeneous deepfake dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9739–9748, 2023.

[24] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[25] R. Roy, I. Joshi, A. Das, and A. Dantcheva. 3d cnn architectures and attention mechanisms for deepfake detection. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pages 213–234. Springer International Publishing Cham, 2022.

[26] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[27] H. Tiwari, V. K. Kurmi, K. Venkatesh, and Y.-S. Chen. Occlusion resistant network for 3d face reconstruction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 813–822, 2022.

[28] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.

[29] J. Xiang and G. Zhu. Joint face detection and facial expression recognition with mtcnn. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 424–427, 2017.

[30] D. Xie, P. Chatterjee, Z. Liu, K. Roy, and E. Kossi. Deepfake detection on publicly available datasets using modified alexnet. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1866–1871, 2020.

[31] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.

[32] Y. Zhou and S.-N. Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021.