FMFCC-V: An Asian Large-Scale Challenging Dataset for DeepFake Detection

Gen Li
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100195, China
ligen1@iie.ac.cn

Xianfeng Zhao*
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100195, China
zhaoxianfeng@iie.ac.cn

Yun Cao
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100195, China
caoyun@iie.ac.cn

Pengfei Pei
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100195, China
peipengfei@iie.ac.cn

Jinchuan Li
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100195, China
lijinchuan@iie.ac.cn

Zeyu Zhang
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100195, China
zhangzeyu@iie.ac.cn

ABSTRACT

The abuse of DeepFake technique has raised enormous public concerns in recent years. Currently, the existing DeepFake datasets suffer some weaknesses of obvious visual artifacts, minimal Asian proportion, backward synthesis methods and short video length. To make up these weaknesses, we have constructed an Asian large-scale challenging DeepFake dataset to enable the training of DeepFake detection models and organized the accompanying video track of the first Fake Media Forensics Challenge of China Society of Image and Graphics (FMFCC-V). The FMFCC-V dataset is by far the first and the largest public available Asian dataset for DeepFake detection, which contains 38102 DeepFake videos and 44290 pristine videos, corresponding more than 23 million frames. The source videos in the FMFCC-V dataset are carefully collected from 83 paid individuals and all of them are Asians. The DeepFake videos are generated by four of the most popular face swapping methods. Extensive perturbations are applied to obtain a more challenging benchmark of higher diversity. The FMFCC-V dataset can lend powerful support to the development of more effective DeepFake detection methods. We contribute a comprehensive evaluation of six representative DeepFake detection methods to demonstrate the level of challenge posed by FMFCC-V dataset. Meanwhile, we provide a detailed analysis of the top submissions from the FMFCC-V competition.

CCS CONCEPTS

• **Security and privacy** → **Social network security and privacy**; • **Information systems** → *Multimedia databases.*

KEYWORDS

DeepFake dataset, DeepFake detection, detection benchmark, forensics competition

ACM Reference Format:
Gen Li, Xianfeng Zhao, Yun Cao, Pengfei Pei, Jinchuan Li, and Zeyu Zhang. 2022. FMFCC-V: An Asian Large-Scale Challenging Dataset for DeepFake Detection. In *Proceedings of the 2022 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '22), June 27–28, 2022, Santa Barbara, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531536.3532946

1 INTRODUCTION

Face swapping has become a disconcerting problem in the last few years. In particular, the emerging of deep learning technology makes face swapping much easier and more convincing. DeepFake is a representative type of face swapping techniques which refers to the set of deep learning based face forgery methods that can create fake portrait videos by swapping the face of a target individual by the face of a source individual, retaining the facial expression and head poses of the target individual [53]. As a result, open software and mobile application have been released opening the door to anyone to create fake portrait videos, without any experience

*Corresponding author.

IH&MMSec '22, June 27–28, 2022, Santa Barbara, CA, USA

ACM ISBN 978-1-4503-9355-3/22/06.
https://doi.org/10.1145/3531536.3532946

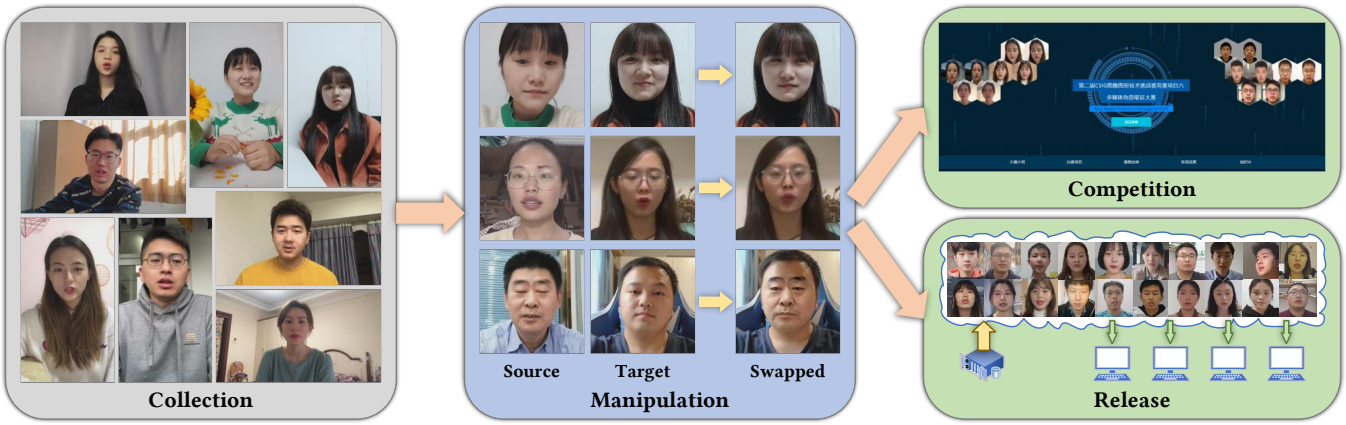

Figure 1: FMFCC-V dataset is an Asian large-scale challenging dataset for DeepFake Detection.

in the field needed [43]. These DeepFake videos, known as DeepFakes, have raised enormous public concerns for their huge risks to create serious political, social, financial and legal consequences [40]. Accordingly, there is a dire need for countermeasures to be in place promptly, particularly innovations that can effectively detect portrait videos that have been manipulated.

One promising countermeasure against DeepFakes is active defense strategy, namely DeepFake detection. Several approaches have been proposed by the researchers to expose DeepFakes. However, training DeepFake detection models generally requires a large amount of both pristine videos and DeepFake videos. The cost of producing the hundreds of thousands of DeepFake videos is often prohibitive. Fortunately, various groups have contributed some public available DeepFake datasets such as the UADFV dataset [56], the DeepFake-TIMIT (DF-TIMIT) dataset [32], the FaceForensics++ (FF++-DF) dataset [46], the Google DeepFake Detection (DFD) dataset [44], the Celeb-DF dataset [37], the DeeperForensics-1.0 (DPF) dataset [29] and the FaceBook DeepFake Detection Challenge (DFDC) dataset [19]. The availability of these datasets has provided essential avenues for research into DeepFake detection. On the other hand, the aforementioned DeepFake video datasets suffer some weaknesses. First, several common visual artifacts such as low quality faces, visible blending boundaries, unmatched color modes and blurry features can be found in videos. These artifacts are likely the result of inadequate training process and imperfect synthesis step. Second, there is a lack of diversity in the aspects of head poses, genders, ages, scenes and lengths in these datasets. It is mainly because of the incomplete plan of collecting source videos. Third, the proportion of Asian videos is minuscule. The vast majority of videos derive from Western which is unfriendly to the DeepFake detection research orientation to Asian.

Due to the above three drawbacks, DeepFake detection models trained on these existing datasets may not fully generalize to wild DeepFakes in the real world. Based on current knowledge and experience, we believe that the effectiveness of DeepFake detection models can only be enhanced when trained with a dataset that is exhaustive enough to encompass as many various DeepFakes as possible. To better support the development and evaluation of more effective DeepFake detection methods, we have constructed a new large-scale challenging Asian DeepFake dataset and organized the accompanying video track of the first Fake Media Forensics Challenge of China Society of Image and Graphics (FMFCC-V). As is graphically depicted in Figure 1, we summarize our two main contributions as follows.

Our first major contribution is the FMFCC-V dataset. The FMFCC-V dataset is by far the first and the largest public available Asian dataset for DeepFake Detection. There are in total 38102 DeepFake videos and 44290 pristine videos, corresponding more than 23 million frames, in the FMFCC-V dataset. The source videos are collected from 83 paid individuals speaking in a variety of conditions for roughly 40 minutes each. All individuals are Asians and give consents to the use and manipulation of their faces by signing a formal agreement. The DeepFake videos are generated using four kinds of most popular face swapping methods for roughly 16 minutes each before post processing. In addition, we introduced diversity into both DeepFake videos and pristine videos through deliberate addition of twelve kinds of perturbations, simulating real world scenarios. Based on the FMFCC-V dataset and other existing DeepFake datasets, we benchmark video-level results of six representative DeepFake detection methods. This is a comprehensive performance evaluation of DeepFake detection methods, offering insights into the current status and future strategy in DeepFake detection. The FMFCC-V dataset can be downloaded from *https://github.com/iiecasligen/FMFCC-V*.

Beyond building and releasing the FMFCC-V dataset, our second major contribution is a completed benchmark competition using this dataset. The FMFCC-V competition attracted over 400 contestants who come from 60 organizations such as Institute of Automation, Chinese Academy of Sciences, Wuhan University, Shenzhen University, University of Science and Technology of China, Zhejiang University, South China University of Technology, National University of Defense Technology and so on. The monetary prizes provided a large incentive for these contestants to dedicate a lot of time and computational resources to optimize DeepFake detection algorithms for benchmarking. Compared with the DFDC competition [3] organized by FaceBook, we only gather the metadata of predictions rather than the source codes and models for protecting the intellectual property rights of the contestants. Based on the

result of the FMFCC-V competition, we provided a detailed analysis of top submissions. The homepage of FMFCC-V competition is *http://fmfcc.net*.

2 RELATED WORK

2.1 DeepFake Generation Methods

Nowadays, there are many sophisticated DeepFake generation methods based on the improved Generative Adversarial Networks (GAN) [24] or Auto-Encoder (AE) [31] have been presented by researchers. Most of these methods have not been in mainstream as available software tools that anyone can use easily. However, several independent open-source face swapping tools and applications such as *DFaker* [4], *FakeApp* [11], *faceswap* [10], *faceswap-GAN* [1] and *DeepFaceLab* [9] have been frequently used to generate DeepFake videos circulated on the Internet and in the existing DeepFake datasets. Most of these DeepFake generation methods adopt the classical AE architecture to synthesize fake face images. An overview of the basic face swapping process based on AE is illustrated in Figure 2.

Before face synthesis, the face images and facial landmarks are extracted from the video frames. The landmarks are used to align face images to a standard configuration. Then, the aligned face images are cropped and fed into an AE to train the face swapping model. The AE is usually formed by encoder-decoder structure. In the training process, the encoder extracts latent features of face images and the decoder is used to reconstruct the face images. To swap identities between target images and source images, there is a need of two encoder-decoder pairs where each pair is used to train on a face image set. Specifically, the parameters of two encoders are shared between two encoder-decoder pairs which enables the common encoder to find and learn the identity-independent attributes between two sets of face images. The two encoder-decoder pairs are trained in an unsupervised manner and optimize their parameters to minimize the reconstruction errors. In the synthesis process, any image containing a target face can be encoded through the shared encoder and decoded with the source decoder. Then, the synthesized faces are warped back to the configuration of the original target faces. The last step is smoothing the boundaries between the synthesized regions and the original video frames. Given a video, the face in each frame can be generated to replace the original face following the above process with little manual intervention.

2.2 Existing DeepFake Datasets

Building a DeepFake dataset for the research of DeepFake detection methods requires a huge amount of effort on data collection and manipulation. Up to now, there are some public available and frequently used DeepFake datasets. Based on release time, synthesis algorithms and data size, these DeepFake datasets can be categorized into three generations [37]. Table 1 lists the basic information of these DeepFake datasets. It should be noted that the number of recompression videos, computer graphics methods, expression swapping methods and other non-learned synthesis methods are not counted.

UADFV: The UADFV dataset [56], released in November 2018, is grouped in the first generation. This dataset contains 49 real videos sourced from *YouTube* and 49 DeepFake videos. The DeepFake

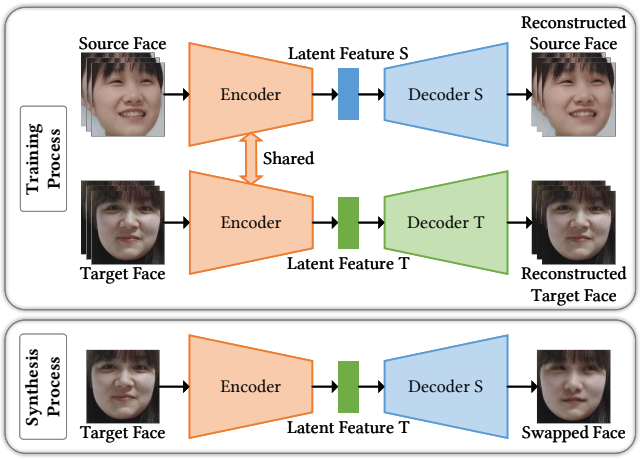

Figure 2: The basic face swapping process.

videos are generated using *FakeApp* [11] application and only one identity is considered in all DeepFake videos. The average length of DeepFake videos is 11 seconds.

DF-TIMIT: The DeepFake-TIMIT dataset [32], released in December 2018, is grouped in the first generation. This dataset contains 320 real videos and 640 DeepFake videos. The real videos come from the VidTIMIT dataset [49]. The DeepFake videos are generated using the public face swapping algorithm *faceswap-GAN* [1]. Regarding the resolution of synthesized face images, two different qualities are considered in DF-TIMIT dataset. They are low quality subset with synthesized faces of 64×64 pixels and high quality subset with synthesized faces of 128×128 pixels. The average length of DeepFake videos is 3 seconds.

FF++-DF: The FaceForensics++ dataset [46], released in January 2019, is grouped in the first generation. This dataset contains 1000 real videos extracted from *YouTube* and 1000 unique DeepFake videos. The DeepFake videos are generated using the public face swapping algorithm *faceswap* [10]. Besides, there are another 3000 fake videos manipulated by computer graphics approach and expression swapping method. Three levels of video quality are considered in FF++-DF dataset. They are raw quality with Constant Rate Factor (CRF) parameter of 0, high quality with CRF parameter of 23 and low quality with CRF parameter of 40. The average length of DeepFake videos is 18 seconds.

DFD: The Google DeepFake Detection dataset [44], released in September 2019, is grouped in the second generation. This dataset contains 363 real videos and 3068 DeepFake videos. The real videos are recorded from 28 consented individuals with various genders, ages and ethnic groups in 16 different scenes. The DeepFake videos are generated using an improved implementation of the face swapping algorithm which are not disclosed. The average length of DeepFake videos is 29 seconds.

Celeb-DF: The Celeb-DF dataset [37], released in November 2019, is grouped in the second generation. This dataset contains 590 real videos and 5639 DeepFake videos. The real videos are chosen from public available *YouTube* videos, corresponding to interviews of 59 celebrities with a diverse distribution in their genders, ages and

Table 1: The basic information of the existing DeepFake datasets.

Dataset	Total Fake Videos	Total Videos	Fake Video Maxlength	Uncontrolled Capture	Informed Consent	Consented Subjects	Asian Proportion	Synthesis Methods	Perturbation Methods
UADFV	49	98	44s	√	×	0	0.0%	1	0
DF-TIMIT	640	960	9s	×	×	0	0.0%	1	0
FF++-DF	4000	5000	72s	√	×	0	0.0%	1	0
DFD	3068	3431	67s	√	√	28	0.0%	1	0
Celeb-DF	5639	6229	24s	√	×	0	5.1%	1	0
DPF	10000	60000	72s	×	√	100	25.0%	1	7
DFDC	104500	123654	11s	√	√	960	0.5%	4	9
FMFCC-V	38102	82392	1165s	√	√	83	100.0%	4	12

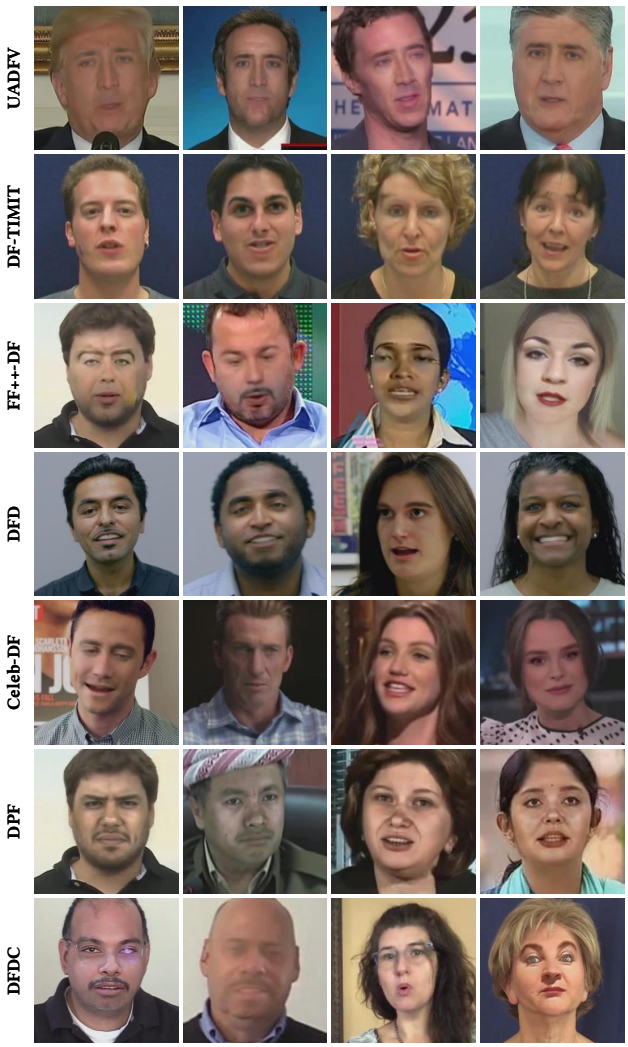

Figure 3: Example faces of the existing DeepFake datasets.

ethnic groups. In particular, about five percent of subjects are Asians. The DeepFake videos are generated using a refined version of a public DeepFake generation algorithm, improving aspects such as the low resolution of the synthesized faces and color inconsistencies. The average length of DeepFake videos is 12 seconds.

DPF: The DeeperForensics-1.0 dataset [29], released in January 2020, is grouped in the third generation. This dataset contains 50000 real videos and 1000 unique DeepFake videos. 100 paid individuals are invited to record the source videos in a controlled studio setting. The participants are selected to ensure variability in genders, ages, skin colors and nationalities. About a quarter of participants belong to the yellow race. All recorded faces are clean without glasses and decorations. The DeepFake videos are produced by a single model and all target individuals come from FF++-DF dataset. Besides, various perturbations are applied on each of the 1000 DeepFake videos for another 9000 augmented DeepFake videos. The average length of DeepFake videos is 18 seconds.

DFDC: The FaceBook DeepFake Detection Challenge dataset [19], released in October 2020, is grouped in the third generation. This dataset contains 19154 real videos and 104500 DeepFake videos. Regarding the data size, it is by far the largest DeepFake dataset. The real videos are recorded from 960 paid individuals with various genders, ages and ethnic groups speaking in a variety of settings. All participants agreed to be filmed, to appear in a machine learning dataset and to have their face images manipulated by machine learning models. The manipulated videos are generated using four deep learning based face swapping methods, two computer graphics methods and an audio swapping method. The average length of DeepFake videos is 10 seconds.

Some examples of the aforementioned DeepFake datasets are shown in Figure 3. During the build process of the FMFCC-V dataset, we invite some paid individuals to record source videos. All participants are Asians with various genders and ages. We get consents from all the participants for using and manipulating their faces. By comparison, the FMFCC-V dataset has advantages in the aspects of the number of DeepFake generation methods, the type of video perturbation methods and the scale of unique DeepFake videos. Therefore, our proposed FMFCC-V dataset is grouped in the third generation.

2.3 DeepFake Detection Methods

DeepFakes are increasingly detrimental to privacy, society security and democracy. Methods for exposing DeepFakes have been proposed as soon as this threat was introduced. In general, The

DeepFake detection methods can be divided into three categories based on the class feature domains.

Methods in the first category are based on the spatial features. Some artifacts, fingerprints or noises may be present in the generated face images or face boundaries. There are some methods exploit visual artifacts [39], inconsistent head poses [56], missing symmetry [33], warping artifacts [36], blurred boundaries [34] and resampling traces [25] to expose manipulated face images. Besides, some detection methods adopt specially designed Deep Neural Networks (DNN) as generic classifiers and let the networks decide which features to analyze. The classical representatives are attention mechanism [59], image reconstruction decoder [41], single classifier [30], capsule network [42] and residual network [52]. Methods in the second category are based on the temporal features. The anomalies between consecutive frames such as optical flow motion [15], eye blinking pattern [35], facial action units [14] and heart rate [23] have been exploited to detect DeepFake videos. Most of temporal features are extracted through Recurrent Neural Networks (RNN) or some variations [47]. Methods in the third category are based on the multi-domain features. This kind of methods exploit multi-branch networks to combine spatial features, temporal features and frequency features for exposing DeepFakes [13, 27, 38, 45, 54]. Besides, the data augmentation technique has been frequently used to enhance the generalization of DeepFake detection methods.

3 THE FMFCC-V DATASET

In order to make up the drawbacks of the existing DeepFake datasets and support the future development of the DeepFake detection methods, we construct the FMFCC-V dataset. The main advantages of FMFCC-V dataset are more generation methods, better visual quality, greater Asian proportion, longer video length and more perturbation types. A comparison of the FMFCC-V dataset with the existing DeepFake datasets is shown as Table 1.

3.1 Source Data

In the stage of source data collection, we invite 83 paid individuals to record the source videos. The average length of source videos is approximate 40 minutes. To avoid the portrait right issues, we obtain consents from all participants for using and manipulating their faces. The statistics of source data are shown in Figure 4. Regarding the genders, the subjects are evenly split between male and female. Or, to be more precise, 42 subjects are males. 41 subjects are females. Regarding the ages, 9.6% subjects are younger than 20. 48.2% subjects are between 20 and 30. 35.0% subjects are between 30 and 40. 7.2% subjects are older than 40. This age distribution is in accordance with the most common age group appearing on the social videos. Regarding the ethnic groups, all subjects are Asians which is a key feature of the FMFCC-V dataset. In addition, the source videos exhibit large range of changes in the aspects of head poses, expressions, backgrounds, resolutions and frame rates. Regarding the head poses, Approximately four fifths of video frames are front face images. The other one fifth of video frames are side face images. Here, the face images are classified as front face when the yaw angle or the pitch angle of the head less than 20 degree, or the face images are classified as side face. Regarding

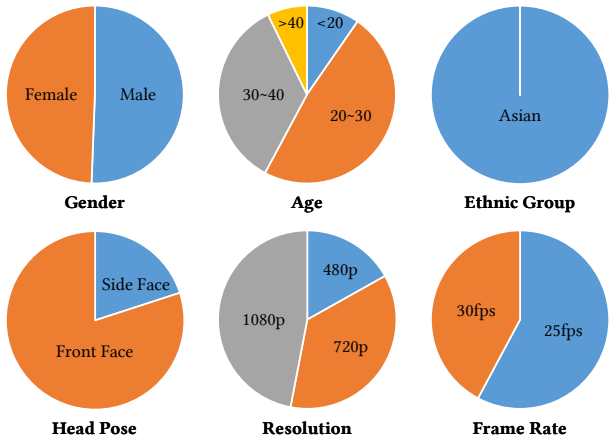

Figure 4: Statistics of our collected source data.

Figure 5: Diversity in genders, ages, head poses and expressions in our collected source data.

the backgrounds, the source videos are recorded in indoor and outdoor settings. Besides, the resolution of source videos mainly includes 480p, 720p and 1080p. The frame rate of source videos mainly includes 25fps and 30fps. Figure 5 shows the diversity in different attributes of our collected source videos. In the end, our collected source data contains over 3700 minutes videos with a total of over 6.3 million frames.

3.2 Synthesis Methods

The DeepFake videos in FMFCC-V dataset are generated using one of four DeepFake synthesis methods of *faceswap* [10], *faceswap-GAN* [1], *DeepFaceLab* [9] and *Recycle-GAN* [16]. These methods are some of the most popular face swapping methods at the time

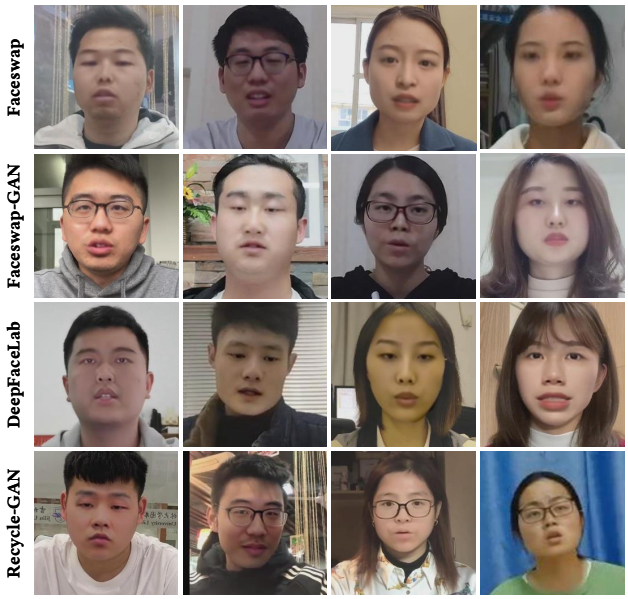

Figure 6: Some cases of the four synthesis methods in the FMFCC-V dataset.

the dataset was created. In the following, take the off-the-shelf software *DeepFaceLab* as an example, we will describe the principle of fake video synthesis in detail. The other three synthesis methods will be described in brief. Some cases of the four synthesis methods in the FMFCC-V dataset are shown in Figure 6.

Faceswap: The *faceswap* [10] method is an early face swapping algorithm based on AE structure. AE structure consists of two encoder-decoder pairs with a shared encoder that are trained to reconstruct face images of the source individual and the target individual, respectively. This structure encourages the shared encoder to learn common features across both identities, while each decoder learns identity-specific features. During the face swapping process, the decoder associated with the source individual takes the encoding of a target face and generate a fake source face. The fake source face is then blended with the rest of the target image using Poisson blending technique. Regarding the target video, the target face in each frame can be generated to replace the original face through the trained face swapping model and post processing steps.

Faceswap-GAN: The *faceswap-GAN* [1] method is an improved vision of the *faceswap* method. Besides the basic structure of AE, the main difference compared to the *faceswap* method is two additional discriminators behind the target decoder and the source decoder. The discriminator learns to distinguish between real face images and reconstructed face images. The success of discriminator is the idea of an classification loss that forces the reconstructed face images to be, in principle, indistinguishable from real face images. This strategy is particularly powerful for image generation task, as this is exactly the objective that much of computer graphics aims to optimize. With the help of the discriminators, the quality of generated face images can be better.

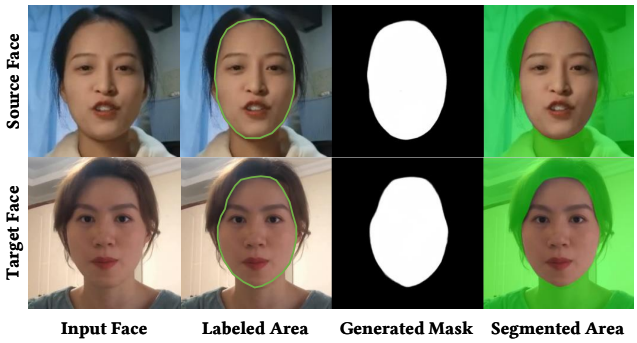

Figure 7: Example segmentation masks generated by face segmentation model.

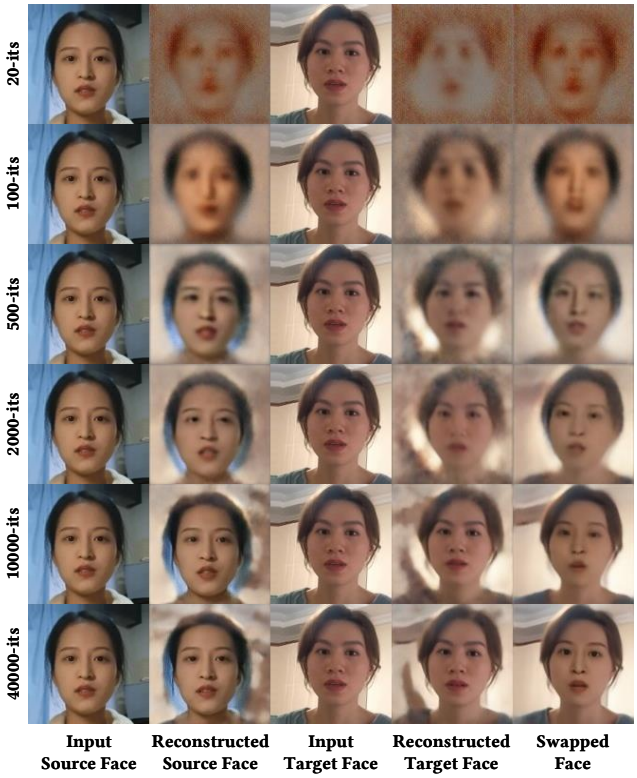

Figure 8: Example results of the processes of face reconstruction and face swapping following the change of the number of training iterations.

DeepFaceLab: The *DeepFaceLab* [9] method is a popular face swapping tool complicated three main phases of extraction, training and conversion. The extraction phase aims to extract the face images from video frames. This phase consists of three processing parts of face detection, face alignment and face segmentation. In the step of face detection, the faces can be extracted by the default face detector. In the step of face alignment, the extracted facial landmarks are used to align faces to a standard template. In the

Table 2: Twelve types of perturbations in FMFCC-V dataset.

No.	Perturbation Type	Apply Ratio
1	Gauss Noise or Multiplicative Noise	0.8
2	Random Brightness	0.8
3	Random Contrast	0.8
4	Gaussian Blur or Motion Blur	0.6
5	Altering Hue Value	0.5
6	Altering Saturation Value	0.5
7	To Sepia or To Gray	0.2
8	Image Compression	0.2
9	Coarse Dropout	0.1
10	Sharpen	0.1
11	Rotate	0.1
12	Horizontal Flip	0.1

step of face segmentation, we use the visual editor to eliminate the occlusion of hands, glasses and any other objects which may cover the face somehow and control specific areas for swapping. Based on the labeled masks, the face segmentation model can be trained to generate fine-grained masks of all video frames. Examples of generated face segmentation mask are shown in Figure 7. After the phase of extraction, everything needs in the training phase is already prepared. Next, we take the LIAE structure as face swapping framework. LIAE structure ia a more complex structure with a shared encoder, two independent inters and a shared decoder. The main difference compared to the traditional AE structure is that the first inter is used to generate both latent codes of source faces and the target faces while the second inter only output the latent codes of the target faces. Then, the latent codes are put into the shared decoder. We can get the reconstructed source faces and target faces alongside with their masks. Figure 9 shows the example results of the processes of face reconstruction and face swapping following the change of the number of training iterations. In the phase of conversion, the first step is to transform the generated face alongside with its mask from the trained decoder to the original position of the target image in source face. The next step is blending and sharpening, with the ambition for the generated face to seamlessly fit with the target image along its outer contour. In a similar way, the target face in each frame can be processed through above steps.

Recycle-GAN: The *Recycle-GAN* [16] method is an unsupervised approach for video retargeting that enables the transfer of sequential content from one domain to another while preserving the style of the target domain. This method combines both spatial and temporal information along with adversarial losses for content translation and style preservation. Move to our face swapping task, the contents of the source video can be transferred to the target video and the generated fake video will be in the style of the target video.

3.3 Date Augmentation

In order to increase the diversity of the FMFCC-V dataset, we apply various perturbations such as adding noise, blurring, darkening, brightening, adding contrast and flipping to each frame to better simulate videos in real world. Specifically, as shown in Table 2,

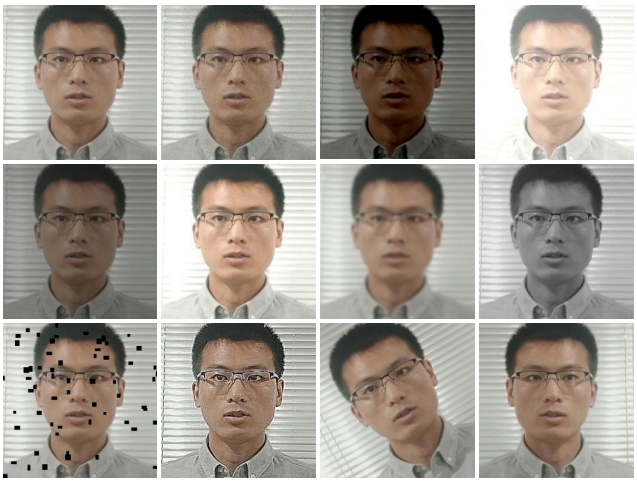

Figure 9: An example face containing various perturbations. From left to right, top row: original, adding noise, darkening, brightening. Middle row: blurring, increasing contrast, reducing contrast, converting to grayscale. Bottom row: dropout parts, sharpening, rotating and horizontal flipping.

twelve types of different perturbations are applied to both DeepFake videos and pristine videos. Each of these perturbations includes different intensity levels. We apply random-type, random-level perturbations to all DeepFake videos and pristine videos, producing an augmented subset. In principle, each augmented videos may be subjected to a mixture of more than one perturbation. See Figure 9 for visual examples of the different perturbations. The variability of perturbations improves the diversity of the FMFCC-V dataset to better imitate the data distribution of the real world.

3.4 Dataset Contents

To meet the requirements of different application scenarios, we provide two versions of the DeepFake dataset. One is the long version of the FMFCC-V dataset and the other is the short version of the FMFCC-V dataset. The main difference between the long version dataset and the short version dataset is that the length of videos in the long version dataset is much greater than in the short version dataset. Besides, the short version dataset has a larger scale and a diverse distribution.

Long Version: The FMFCC-V long version dataset consists of 83 pristine videos and 192 DeepFake videos, corresponding more than 7.7 million frames. Each pristine video corresponds to one unique subject. The DeepFake videos are generated using one of the four DeepFake synthesis methods mentioned above. All target videos and source videos used in the DeepFake generation process come from these 83 pristine videos. The length of all videos in the long version dataset is approximate 16 minutes. No perturbations are applied to this dataset.

Short Version: The FMFCC-V short version dataset consists of 44290 pristine videos and 38102 DeepFake videos, corresponding more than 23 million frames. The length of all videos in the

short version dataset is approximate 10 seconds. The perturbations mentioned above are applied to half of these pristine videos and DeepFake videos. Those disturbed videos formed a augmentation subset which can be used to evaluate the robust performance of the DeepFake detection methods.

4 DETECTION BENCHMARK

In order to survey the challenge level of the existing DeepFake datasets and our constructed FMFCC-V dataset, we perform a comprehensive performance evaluation of DeepFake detection, with the six of the current representative DeepFake detection methods.

4.1 Compared Methods

We consider six DeepFake detection methods in our evaluation experiments. Those methods are either robust in various detection scenarios or classic in the field of video forensics. All of those DeepFake detection methods have public available testing codes and corresponding models.

XceptionNet: The Method in [8, 46] adopts Xception network [17] as the backbone architecture for DeepFake classification. XceptionNet is a traditional Convolutional Neural Network (CNN) based on separable convolutions with residual connections. In the task of DeepFake detection, the final fully connected layer of XceptionNet is replaced with two outputs. The other layers stay the same. The XceptionNet model used in our evaluation experiments is trained on the FF++-DF dataset.

CapsuleNet: The Method in [2, 42] uses Capsule network [48] based on VGG-19 network [50] as the backbone architecture. This method consists of three primary capsules and two output capsules. The latent features extracted by part of the VGG-19 network are the inputs, which are distributed to the three primary capsules. The output features are dynamically routed to the two output capsules. The CapsuleNet model used in our evaluation experiments is trained on the FF++-DF dataset.

DSP-FWA: The Method in [5, 36] employs a dual spatial pyramid strategy on both image and feature level to expose the face warping artifacts introduced by the resizing and interpolation operations in the basic DeepFake generation algorithms. The spatial pyramid pooling module [26] in this method can be used to better handle the variations in the resolutions of the face images. The DSP-FWA model used in our evaluation experiments is trained on unpublished DeepFake dataset collected by the authors.

Heavy-DA: The Method in [12] uses EfficientNet-B7 [51] pretrained with ImageNet and noisy student [55] as the backbone architecture. The heavy augmentations such as adding noise, blurring, rotating, altering contrast, darkening and brightening are applied to training data. Besides, dropping structured parts of faces during training is also used as a form of augmentation. The Heavy-DA model used in our evaluation experiments is trained on the DFDC dataset.

WS-DA: The Method in [7] ensembles two Weakly Supervised Data Augmentation Network (WS-DAN) models [28] and a XceptionNet classifier to produce predictions. The two WS-DAN models adopt EfficientNet-B3 and XceptionNet as feature extractors, respectively. The XceptionNet classifier used the original design with two-class output and applied some augmentations to improve model generalization. The WS-DA model used in our evaluation experiments is trained on the DFDC dataset.

Mixup-DA: The Method in [6] consists of three EfficientNet-B7 models. One model runs on frame sequences where the three-dimensional convolution are added to each block of the EfficientNet-B7 structure. The other two models work frame-by-frame and differ in the size of the face crop and augmentations during training. The mixup technique [57] is applied on aligned real-fake face pairs to tackle overfitting problem. The Mixup-DA model used in our evaluation experiments is trained on the DFDC dataset.

4.2 Experimental Settings

We evaluate the overall detection performance using the video-level Area Under ROC Curve (AUC) score for the reason that fake videos are much more menacing than manipulated images. To balance the computational precision and computational complexity, 100 DeepFake videos and 100 pristine videos were picked out from each DeepFake dataset randomly for constructing the testing dataset in our evaluation experiments. Specially, we take all the videos in the UADFV dataset as the testing dataset for the reason that the number of videos in the UADFV dataset is smaller than 100. The testing videos of the DF-TIMIT dataset come from the high quality subset. The testing videos of the FF++-DF dataset come from the raw quality subset. For each video, we perform evaluations only on one fifth of all frames. There is no significant difference in the number of sampled frames of testing videos for the reason that there is no significant difference in the length of testing videos. The prediction of each video is the average value of all sampled frames. We evaluate performance of each detection method using the inference codes and the published detection models with the default settings of the hyper-parameter. The average detection performance is presented as an indicator of the challenge levels of various DeepFake datasets.

4.3 Results and Analysis

After carried out all evaluation experiments, we present a comprehensive DeepFake detection benchmark at video level. Table 3 and Figure 10 show AUC scores and ROC curves of the compared DeepFake detection methods on the existing DeepFake datasets.

According to the average AUC scores of the compared DeepFake detection methods on each dataset, in general, the FMFCC-V dataset is a challenging dataset to the current DeepFake detection methods. The overall performance of the compared DeepFake detection methods on FMFCC-V dataset is lower than most of the existing DeepFake datasets. To be exact, the average AUC score of the FMFCC-V dataset is reduced 0.1409, 0.0010, 0.1644, 0.0129 and 0.0047 compared with the UADFV dataset, the DF-TIMIT dataset, the FF++-DF dataset, the Celeb-DF dataset and the DPF dataset, respectively. On the other hand, The average AUC score of the FMFCC-V dataset is increased 0.0281 and 0.0110 compared with the DFD dataset and the DFDC dataset which indicates that the challenge levels of the DFD dataset and the DFDC dataset are marginally higher than the FMFCC-V dataset. But, the data size of the FMFCC-V dataset is much greater than the DFD dataset. The Asian proportion and the video mixlength of the FMFCC-V dataset is much greater than the DFDC dataset. According to the performance of the DSP-FWA method, the visual quality of the videos in the FMFCC-V dataset

Table 3: AUC scores of the compared DeepFake detection methods on the existing DeepFake datasets.

Method↓ Dataset→	UADFV	DF-TIMIT	FF++-DF	DFD	Celeb-DF	DPF	DFDC	FMFCC-V
XceptionNet [8, 46]	0.9742	0.5000	1.0000	0.4595	0.5096	0.4604	0.5488	0.4741
CapsuleNet [2, 42]	0.8661	0.7418	0.9998	0.7213	0.7641	0.5877	0.6479	0.7162
DSP-FWA [5, 36]	0.9500	0.9978	0.9004	0.6123	0.7245	0.9398	0.6158	0.7993
Heavy-DA [12]	0.9642	0.8668	0.9744	0.9810	0.9851	0.9814	0.9992	0.9726
WS-DA [7]	0.9971	0.9454	1.0000	0.9464	0.9871	0.9764	0.9996	0.9535
Mixup-DA [6]	0.9708	0.8313	0.9889	0.9875	0.9836	0.9595	0.9993	0.9613
Average	0.9537	0.8138	0.9772	0.7847	0.8257	0.8175	0.8018	0.8128

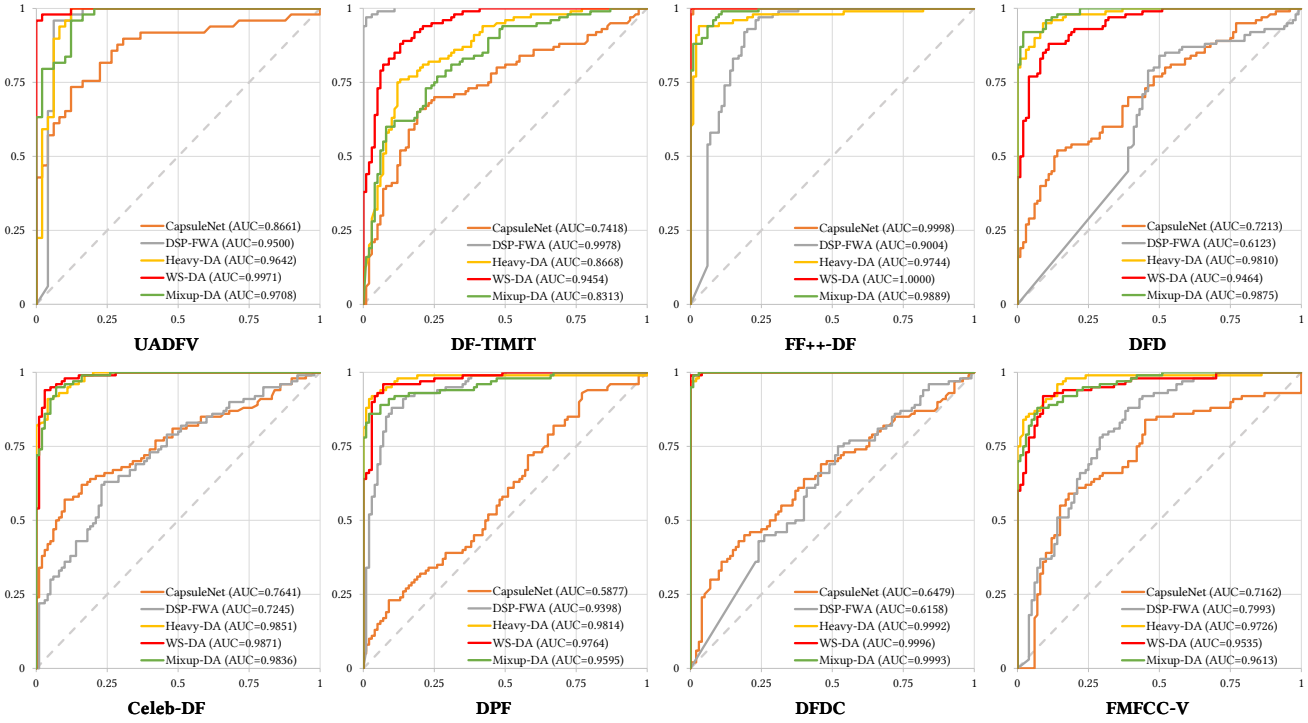

Figure 10: ROC curves of top five DeepFake detection methods on the existing DeepFake datasets.

has been improved obviously compared with the videos in the first generation datasets. Furthermore, the difficulty level for detection of the FMFCC-V dataset is higher than the other datasets to the data augmentation based methods mentioned above.

5 THE FMFCC-V COMPETITION

The goal of the FMFCC-V competition is to spur researchers to build innovative new technologies that can help to expose DeepFakes. Based on the released FMFCC-V dataset, contestants just need to submit the metadata of predictions into the FMFCC-V competition platform for ranking the performance of their solutions. Submission of source codes and models are not required for protecting the intellectual property rights of contestants. All submissions are evaluated in the same way. Results are shown on the leaderboard in real time.

5.1 Evaluation Metric

For the purpose of ranking submissions, a threshold-free evaluation metric, LogLoss, is used to estimate the performance of predictions. The function of LogLoss can be written as:

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \ln\left(p_i\right) + \left(1-y_i\right)\ln\left(1-p_i\right)\right], \qquad (1)$$

where N is the number of videos being predicted. p_i is the predicted probability of the video being fake. y_i is 1 if the video is fake. y_i is 0 if the video is real. In summary, a smaller LogLoss is better. The use of the logarithm provides extreme punishments for being both confident and wrong. In the worst possible case, a prediction that a video is true when it is actually false will add infinite to error score. In order to avoid this, predictions are bounded away from the extremes by a small value, 10^{-12}, in our evaluation metric.

Table 4: Detection metrics of the top ten submissions in the preliminary phase and the final phase.

Team Name	Preliminary Phase						Final Phase					
	Rank	LogLoss	AUC	PRE	REC	ACC	Rank	LogLoss	AUC	PRE	REC	ACC
WLAR	9	0.1246	0.9901	0.9487	0.9429	0.9514	1	0.1265	0.9908	0.9431	0.9613	0.9517
CCDS	4	0.0922	0.9949	0.9638	0.9721	0.9710	2	0.2563	0.9598	0.8875	0.8787	0.8837
Tom For Jerry	7	0.1185	0.9914	0.9580	0.9419	0.9553	3	0.2646	0.9589	0.8777	0.8853	0.8810
RealDeepfake-SZU	8	0.1222	0.9906	0.9498	0.9479	0.9540	4	0.2883	0.9539	0.8816	0.8633	0.8738
ChuanJuDaShi	1	0.0406	0.9987	0.9880	0.9824	0.9868	5	0.3098	0.9495	0.8397	0.8660	0.8503
prml	10	0.1268	0.9894	0.9528	0.9394	0.9518	6	0.3175	0.9495	0.8511	0.8767	0.8617
NCC	5	0.1021	0.9952	0.9811	0.9467	0.9678	7	0.3206	0.9482	0.8724	0.8520	0.8637
Maya	2	0.0772	0.9964	0.9636	0.9908	0.9790	8	0.3612	0.9470	0.8857	0.8887	0.8870
SCUT	6	0.1047	0.9911	0.9711	0.9471	0.9635	9	0.3838	0.9376	0.8651	0.8120	0.8427
SCUT 414	3	0.0904	0.9853	0.9729	0.9760	0.9770	10	0.4809	0.8872	0.8353	0.8047	0.8230
Average	-	0.0999	0.9923	0.9650	0.9587	0.9657	-	0.3110	0.9482	0.8739	0.8689	0.8718

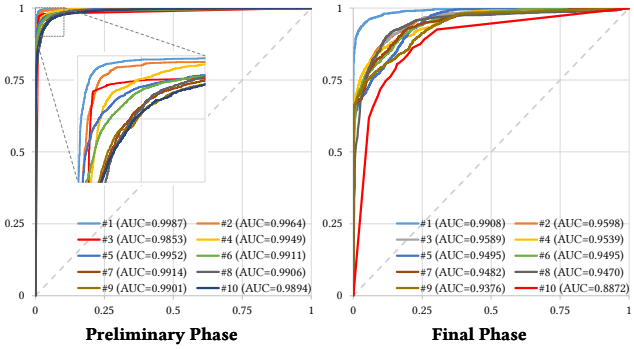

Figure 11: ROC curves of the top ten submissions in the preliminary phase and the final phase.

5.2 Competition Schedule

The FMFCC-V competition consists of the preliminary phase and the final phase. Only top ten teams of the preliminary phase can reach the final phase.

Preliminary Phase: In the preliminary phase, we released two datasets of the public training dataset and the public testing dataset. The public training dataset, containing labels for the videos, was available for contestants to train their DeepFake detection models. This dataset consists of 10000 ten second video clips, in which 40% (4000 clips) included DeepFake videos and 60% (6000 clips) included pristine videos. It was allowed that contestants train their models with the other external DeepFake datasets. The public testing dataset is what competition platform computes the preliminary leaderboard against. This dataset consists of 20000 ten second video clips, in which 45% (9000 clips) included DeepFake videos and 55% (11000 clips) included pristine videos. The two datasets were constructed from undisturbed videos of the FMFCC-V short version dataset. About 50 unique subjects were used in this phase. Contestants can run their detection model in their own environments against the public testing dataset. The metadata of predictions were submitted into the FMFCC-V competition platform. The scores of all teams were posted to the preliminary leaderboard of the FMFCC-V competition platform in real time. Contestants were limited to three submissions every day.

Final Phase: In the final phase, we constructed a private testing dataset which was privately held outside the FMFCC-V competition platform, and was used to compute the final leaderboard. This dataset consists of 3000 ten second video clips, in which 50% (1500 clips) included DeepFake videos and the other 50% (1500 clips) included pristine videos. This dataset was constructed from disturbed videos of the FMFCC-V short version dataset. About 30 unique subjects were used in this phase and none of which were a part of the released datasets in the preliminary phase. Unlike the public testing dataset, a few videos of the private testing dataset included organic content found in the wild scenario. After the preliminary deadline, top ten teams were invited to deploy their algorithms on the FMFCC-V competition platform remotely. Then, we can run their algorithms on a pair of NVIDIA RTX 3090 GPU against the private testing dataset. The prediction submissions were returned to the FMFCC-V competition platform for computing their final leaderboard scores. Results had to run over all videos in the private testing dataset within two hours.

5.3 Results and Analysis

After the FMFCC-V competition ended, 196 teams submitted the results in the preliminary phase in total and top ten teams reached the final phase. In order to appraise the performance of the submissions from many aspects, Besides the LogLoss, we also calculated the detection metrics of AUC score, Precision (PRE), Recall (REC) and Accuracy (ACC). The functions of PRE, REC and ACC can be written as:

$$PRE = \frac{N_{TP}}{N_{TP} + N_{FP}}, \tag{2}$$

$$REC = \frac{N_{TP}}{N_{TP} + N_{FN}}, \tag{3}$$

$$ACC = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{TN} + N_{FN}}, \tag{4}$$

where N_{TP}, N_{FP}, N_{TN} and N_{FN} signify the number of true positive videos, false positive videos, true negative videos and false negative videos. Here, we set the classification threshold to 0.5. Finally,

shown in Table 4 are detection metrics of the top ten submissions in the preliminary phase and the final phase. Figure 11 shows ROC curves of the top ten submissions in the preliminary phase and the final phase, respectively.

As expected, the best submissions achieved very good detection performance on the public testing datset in the preliminary phase for the reason that the videos in the public training dataset has the same distribution with the public testing dataset. However, the detection performance of most of the top ten submissions in the final phase is severely degraded for the reason that the videos in the private testing dataset were added various perturbations, none of which were seen in the public training dataset and public testing dataset. Specially, *WLAR* ranked the ninth place with the LogLoss of 0.1246 and ACC of 0.9514 in the preliminary phase, achieved the first place with the LogLoss of 0.1265 and ACC of 0.9517 in the final phase. *ChuanJuDaShi* ranked the first place with the LogLoss of 0.0406 and ACC of 0.9868 in the preliminary phase, achieved the fifth place with the LogLoss of 0.3098 and ACC of 0.8503 in the final phase. The average LogLoss in the final phase are increased 0.2111 compared with the average LogLoss in the preliminary phase. The average ACC in the final phase are reduced 0.0939 compared with the average ACC in the preliminary phase. The competition results illustrate the challenge of the FMFCC-V dataset for DeepFake detection methods.

5.4 Winning Solutions

Here, we provide a brief description of the top three winning solutions of the final phase.

The first solution, *WLAR*, uses MTCNN [58] for face detection. The backbone architecture consists of a face classification model and a face clustering model. The face classification model based on EfficientNet-B7 [51] is used for predicting the manipulation probability of single frame. The face clustering model is used for individual clustering. Then, the manipulation probability of each video is the fusion result of all clustered individuals. The second solution, *CCDS*, uses two-pathway structure for deep feature extraction. One pathway is browsing pathway which is used for checking the temporal consistency of each video. Another pathway is scrutinizing pathway which is used for predicting the manipulation probability of each key frame. The backbone network includes three mainstream networks of SlowFastNet [22], X3DNet [21] and MVitNet [20]. The third solution, *TomForJerry*, uses MTCNN and RetinaFace [18] for face detection in different detection stages. The backbone network is EfficientNet-B0. Multiple negative sample generation methods are used for avoiding overfitting during detection model training process.

6 CONCLUSION

In this paper, we proposed an Asian large-scale challenging dataset FMFCC-V for DeepFake detection. FMFCC-V dataset consists of 38102 DeepFake videos and 44290 pristine videos. The pristine videos are recorded from 83 paid Asians. All individuals give consents to the use and manipulation of their faces. The DeepFake videos are generated by four of the most popular DeepFake synthesis methods. Twelve kinds of perturbations are applied to both DeepFake videos and pristine videos. Compared with the existing DeepFake datasets, FMFCC-V dataset contains more diverse DeepFake generation methods, better visual quality, greater Asian proportion, longer video length and more perturbation types. FMFCC-V dataset can serve as a useful supplementary to the existing DeepFake datasets to support the development and evaluation of more effective DeepFake detection methods against DeepFakes in real world. We conducted a comprehensive evaluation of six representative DeepFake detection methods on the existing DeepFake datasets and FMFCC-V dataset. The experimental results indicate that FMFCC-V dataset is indeed a challenging DeepFake dataset. Based on our constructed FMFCC-V dataset, we have organized the FMFCC-V competition which attracted about 400 contestants. We provided a detailed analysis of top submissions.

ACKNOWLEDGMENTS

This work was supported by National Key Technology Reseach and Development Program under 2019QY2202 and 2020AAA0140000.

REFERENCES

[1] 2019. *faceswap-GAN.* Retrieved October 4, 2019 from https://github.com/shaoanlu/faceswap-GAN
[2] 2020. *CapsuleNet.* Retrieved December 15, 2020 from https://github.com/nii-yamagishilab/Capsule-Forensics-v2
[3] 2020. *Deepfake Detection Challenge.* Retrieved March 3, 2020 from https://www.kaggle.com/c/deepfake-detection-challenge
[4] 2020. *DFaker.* Retrieved June 19, 2020 from https://github.com/dfaker/df
[5] 2020. *DSP-FWA.* Retrieved June 25, 2020 from https://github.com/yuezunli/DSP-FWA
[6] 2020. *Mixup-DA.* Retrieved June 9, 2020 from https://github.com/NTech-Lab/deepfake-detection-challenge
[7] 2020. *WS-DA.* Retrieved June 14, 2020 from https://github.com/cuihaoleo/kaggle-dfdc
[8] 2020. *XceptionNet.* Retrieved July 15, 2020 from https://github.com/ondyari/FaceForensics
[9] 2021. *DeepFaceLab.* Retrieved November 20, 2021 from https://github.com/iperov/DeepFaceLab
[10] 2021. *faceswap.* Retrieved December 5, 2021 from https://github.com/deepfakes/faceswap
[11] 2021. *FakeApp.* Retrieved October 2, 2021 from https://www.malavida.com/en/soft/fakeapp/
[12] 2021. *Heavy-DA.* Retrieved November 9, 2021 from https://github.com/selimsef/dfdc_deepfake_challenge
[13] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security*. IEEE, 1–7. https://doi.org/10.1109/WIFS.2018.8630761
[14] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. Protecting World Leaders Against Deep Fakes. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Computer Vision Foundation / IEEE, 38–45.
[15] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. 2019. Deepfake Video Detection through Optical Flow Based CNN. In *2019 IEEE/CVF International Conference on Computer Vision Workshops*. IEEE, 1205–1207. https://doi.org/10.1109/ICCVW.2019.00152
[16] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. Recycle-GAN: Unsupervised Video Retargeting. In *Computer Vision - ECCV 2018 - 15th European Conference (Lecture Notes in Computer Science, Vol. 11209)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, 122–138. https://doi.org/10.1007/978-3-030-01228-1_8
[17] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 1800–1807. https://doi.org/10.1109/CVPR.2017.195
[18] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. *CoRR* abs/1905.00641 (2019). arXiv:1905.00641 http://arxiv.org/abs/1905.00641
[19] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. 2020. The DeepFake Detection Challenge Dataset. *CoRR* abs/2006.07397 (2020). arXiv:2006.07397 https://arxiv.org/abs/2006.07397

[20] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale Vision Transformers. *CoRR* abs/2104.11227 (2021). arXiv:2104.11227 https://arxiv.org/abs/2104.11227

[21] Christoph Feichtenhofer. 2020. X3D: Expanding Architectures for Efficient Video Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 200–210. https://doi.org/10.1109/CVPR42600.2020.00028

[22] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision*. IEEE, 6201–6210. https://doi.org/10.1109/ICCV.2019.00630

[23] Steven Lawrence Fernandes, Sunny Raj, Eddy Ortiz, Iustina Vintila, Margaret Salter, Gordana Urosevic, and Sumit Kumar Jha. 2019. Predicting Heart Rate Variations of Deepfake Videos using Neural ODE. In *2019 IEEE/CVF International Conference on Computer Vision Workshops*. IEEE, 1721–1729. https://doi.org/10.1109/ICCVW.2019.00213

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 2672–2680.

[25] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2020. DeepFake Detection by Analyzing Convolutional Traces. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 2841–2850. https://doi.org/10.1109/CVPRW50498.2020.00341

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 9 (2015), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824

[27] Juan Hu, Xin Liao, Wei Wang, and Zheng Qin. 2021. Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network. *IEEE Transactions on Circuits and Systems for Video Technology* (2021), 1–1. https://doi.org/10.1109/TCSVT.2021.3074259

[28] Tao Hu and Honggang Qi. 2019. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification. *CoRR* abs/1901.09891 (2019). arXiv:1901.09891 http://arxiv.org/abs/1901.09891

[29] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 2886–2895. https://doi.org/10.1109/CVPR42600.2020.00296

[30] Hasam Khalid and Simon S. Woo. 2020. OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 2794–2803. https://doi.org/10.1109/CVPRW50498.2020.00336

[31] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6114

[32] Pavel Korshunov and Sébastien Marcel. 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. *CoRR* abs/1812.08685 (2018). arXiv:1812.08685 http://arxiv.org/abs/1812.08685

[33] Gen Li, Yun Cao, and Xianfeng Zhao. 2021. Exploiting Facial Symmetry to Expose Deepfakes. In *2021 IEEE International Conference on Image Processing (ICIP)*. 3587–3591. https://doi.org/10.1109/ICIP42928.2021.9506272

[34] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face X-Ray for More General Face Forgery Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition,*. Computer Vision Foundation / IEEE, 5000–5009. https://doi.org/10.1109/CVPR42600.2020.00505

[35] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. 2018. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In *2018 IEEE International Workshop on Information Forensics and Security*. IEEE, 1–7. https://doi.org/10.1109/WIFS.2018.8630787

[36] Yuezun Li and Siwei Lyu. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Computer Vision Foundation / IEEE, 46–52.

[37] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 3204–3213. https://doi.org/10.1109/CVPR42600.2020.00327

[38] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. 2020. Two-Branch Recurrent Network for Isolating Deepfakes in Videos. In *Computer Vision - ECCV 2020 - 16th European Conference (Lecture Notes in Computer Science, Vol. 12352)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 667–684. https://doi.org/10.1007/978-3-030-58571-6_39

[39] Falko Matern, Christian Riess, and Marc Stamminger. 2019. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In *IEEE Winter Applications of Computer Vision Workshops*. IEEE, 83–92. https://doi.org/10.1109/WACVW.2019.00020

[40] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1 (2021), 7:1–7:41. https://doi.org/10.1145/3425780

[41] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos. In *10th IEEE International Conference on Biometrics Theory, Applications and Systems*. IEEE, 1–8. https://doi.org/10.1109/BTAS46853.2019.9185974

[42] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2307–2311. https://doi.org/10.1109/ICASSP.2019.8682602

[43] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. 2019. Deep Learning for Deepfakes Creation and Detection. *CoRR* abs/1909.11573 (2019). arXiv:1909.11573 http://arxiv.org/abs/1909.11573

[44] Google Research Nick Dufour and Jigsaw Andrew Gully. 2019. *Contributing Data to Deepfake Detection Research*. Retrieved September 24, 2019 from https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[45] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *Computer Vision - ECCV 2020 - 16th European Conference (Lecture Notes in Computer Science, Vol. 12357)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 86–103. https://doi.org/10.1007/978-3-030-58610-2_6

[46] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision*. IEEE, 1–11. https://doi.org/10.1109/ICCV.2019.00009

[47] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Computer Vision Foundation / IEEE, 80–87.

[48] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic Routing Between Capsules. In *Advances in Neural Information Processing Systems*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 3856–3866.

[49] Conrad Sanderson and Brian C. Lovell. 2009. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. In *Advances in Biometrics, Third International Conference (Lecture Notes in Computer Science, Vol. 5558)*, Massimo Tistarelli and Mark S. Nixon (Eds.). Springer, 199–208. https://doi.org/10.1007/978-3-642-01793-3_21

[50] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.1556

[51] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 6105–6114. http://proceedings.mlr.press/v97/tan19a.html

[52] Rubén Tolosana, Rubén Vera-Rodríguez, Julian Fiérrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Inf. Fusion* 64 (2020), 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

[53] Luisa Verdoliva. 2020. Media Forensics and DeepFakes: An Overview. *IEEE J. Sel. Top. Signal Process.* 14, 5 (2020), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101

[54] Xi Wu, Zhen Xie, YuTao Gao, and Yu Xiao. 2020. SSTNet: Detecting Manipulated Faces Through Spatial, Steganalysis and Temporal Features. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2952–2956. https://doi.org/10.1109/ICASSP40776.2020.9053969

[55] Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. 2020. Self-Training With Noisy Student Improves ImageNet Classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 10684–10695. https://doi.org/10.1109/CVPR42600.2020.01070

[56] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8261–8265. https://doi.org/10.1109/ICASSP.2019.8683164

[57] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *6th International Conference on Learning Representations*. OpenReview.net. https://openreview.net/forum?id=r1Ddp1-Rb

[58] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* 23, 10 (2016), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

[59] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-Attentional Deepfake Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 2185–2194.