KoDF: A Large-scale Korean DeepFake Detection Dataset

Patrick Kwon[*]   Jaeseong You[*]   Gyuhyeon Nam   Sungwoo Park   Gyeongsu Chae

MoneyBrain Inc.

Seoul, Republic of Korea

{patrick jaeseongyou ngh3053 daniel gc}@moneybrain.ai

Abstract

A variety of effective face-swap and face-reenactment methods have been publicized in recent years, democratizing the face synthesis technology to a great extent. Videos generated as such have come to be called deepfakes with a negative connotation, for various social problems they have caused. Facing the emerging threat of deepfakes, we have built the Korean DeepFake Detection Dataset (KoDF), a large-scale collection of synthesized and real videos focused on Korean subjects. In this paper, we provide a detailed description of methods used to construct the dataset, experimentally show the discrepancy between the distributions of KoDF and existing deepfake detection datasets, and underline the importance of using multiple datasets for real-world generalization. KoDF is publicly available at https://moneybrain-research.github.io/kodf *in its entirety (i.e. real clips, synthesized clips, clips with adversarial attack, and metadata).*
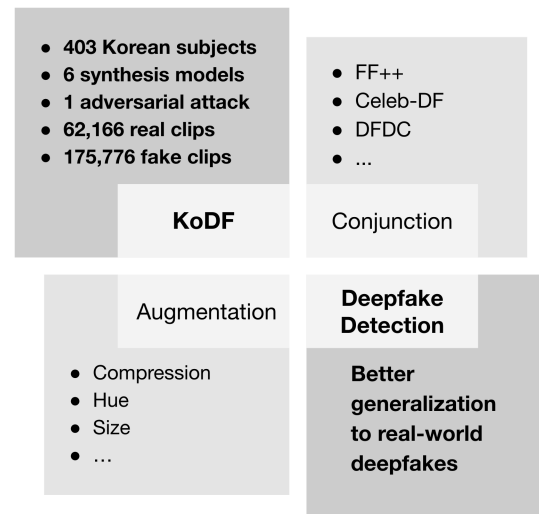
Figure 1. KoDF is a distribution-controlled large-scale Korean deepfake detection dataset aimed to complement other datasets and to accommodate elaborate augmentation techniques for better generalization to real-world deepfakes.

1. Introduction

In recent years, the fabrication of facial content in images and videos has become considerably easier and faster, which previously required heavy computing resources and expert knowledge. Latest deep-learning-based technologies have made it possible to handily produce photorealistic fake images and videos by manipulating facial expressions or swapping faces. Soon the word *deepfake* became the de facto term to indicate such facial forgeries synthesized by deep learning models.

While mostly utilized for innocuous purposes such as parody videos [3] and entertaining apps [2, 13], well-designed deepfakes could be used maliciously to defame an individual [20, 31], propagate disinformation [55, 18], or commit fraud [51, 56]. Due to the growing concerns over deepfakes, there is a recent surge of interest in developing deepfake detection models, and to this end, var-

ious public datasets [37, 48, 39, 25, 36, 26, 23] and benchmarks [8, 6, 4, 7] have been constructed. They have greatly contributed to encouraging, facilitating, and standardizing deepfake detection research.

In line with these efforts, we release the Korean Deep-Fake Detection Dataset (KoDF) featuring important differentiations from the previous deepfake detection datasets. KoDF is the largest among publicly available deepfake detection datasets, containing 175,776 fake clips and 62,166 real clips of 403 subjects. The deepfake samples are generated with six different synthesis models. To counterbalance the Asian demographics underrepresented in the existing deepfake detection databases, the participants of KoDF are comprised mostly of Koreans. Finally, the dataset takes various measures to better manage the data distribution regarding the participants' age, sex, and content. Table 1 compares KoDF with other public deepfake detection datasets in various aspects.

---

[*]Equal contribution

Dataset	Real videos	Fake videos	Total videos	Rights cleared	Agreeing subjects	Total subjects	Methods
UADFV [58]	49	49	98	No	0	49	1
DeepfakeTIMIT [37]	640	320	960	No	0	32	2
FF++ [48]	1,000	4,000	5,000	No	0	N/A	4
Celeb-DF [39]	590	5,639	6,229	No	0	59	1
GDFD [25]	363	3,068	3,431	Yes	28	28	5
DF-1.0 [36]	50,000	10,000	60,000	No[1]	100	100	1
DFDC [23]	23,654	104,500	128,154	Yes	960	960	8
KoDF	62,166	175,776	237,942	Yes	403	403	6

Table 1. Quantitative comparison of KoDF to existing public deepfake detection datasets.

Our contributions are twofold: (1) We propose KoDF that is the largest public deepfake detection dataset, planned and examined for the quality and diversity of its samples. (2) We experimentally demonstrate that none of the established deepfake detection datasets single-handedly suffices in approximating the true deepfake distribution. We then show how utilizing KoDF in conjunction with them for training enhances the generality of a detection model, offering insights into the future strategy in deepfake detection.

2. Related Works

The early deepfake detection databases—the UADFV dataset [58] and the DeepfakeTIMIT dataset [37]—before FaceForensics++ (FF++) [48] and the DeepFake Detection Challenge (DFDC) dataset [23] have limitations in quantity and quality. The number of featured identities does not exceed 50, and the total amount of the real and fake videos are less than 1,000. They are collected from untraceable sources, or the agreements from the subjects regarding possible modification and public use of their face are unclear. The fake clips contain a large number of unrealistic synthesized results. Furthermore, the number of employed synthesis methods is only one or two, failing to capture the diversity in the modern means of facial forgery. The two milestone datasets, FF++ and the DFDC dataset, however, overcome many of these difficulties, and other viable deepfake detection databases have been contributed as well.

FF++ [48] is the first large-scale dataset to contain 1,000 real videos from YouTube and 4,000 fake videos synthesized by two computer-graphics-based and two learning-based methods. Each of 1,000 raw videos is processed through the four chosen methods, resulting in a total of 5,000 clips. Accompanied is a public leaderboard [8] where a deepfake detection model can be evaluated against a hidden test set according to benchmark scenario. Before the release of the DFDC dataset, FF++ served as the de facto

standard deepfake dataset, thus utilized in various research projects [16, 49, 17, 57]. However, it fails to fully address some of the aforementioned issues; the size and the diversity of the dataset are still insufficient for optimal training of high-performing neural architectures comprised of huge numbers of parameters, and the subjects' permissions to appear in the database are missing.

In 2020, Amazon Web Services, Facebook, Microsoft, the Partnership on AI's Media Integrity Steering Committee, and academics have collaborated for DFDC, a large-scale project consisting of a competition, a dataset, and accompanying papers [24, 23]. The DFDC dataset is released as a part of the challenge. It is thus far the second-largest public deepfake dataset next to KoDF, containing over 960 subjects and more than 120,000 videos. To guarantee the variety of the database, the raw clips are taken from different environmental settings, and the synthesized clips are generated by eight different methods. The challenge was highly successful, encouraging a wide range of researchers to partake in developing effective deepfake detection models, followed by an increased number of research publications on the topic of deepfake [32, 38, 40, 53]. However, the DFDC dataset is not without its flaws. Due to the unguided recording process where the participants record themselves, extreme light, audio, and angle conditions are incorporated (e.g. a person talking in a completely dark room), and the data format is inconsistent (e.g. resolution and duration varying across clips). The distribution of participants is not controlled according to age, sex, or race.

Recent years have witnessed other notable public deepfake detection databases varied in focus, composition, and size. Celeb-DF [39] consists of 590 real videos and 5,639 fake videos. The real videos are taken from YouTube, of which the contents are interviews of 59 celebrities. The fake videos are synthesized by an improved face swap method. The Google DeepFake Detection (GDFD) dataset [25] incorporates 3,068 deepfake videos generated based on 363 original videos of 28 consented individuals in 16 different scenes. DeeperForensics-1.0 (DF-1.0) [36] is yet another recent deepfake detection dataset. Its source videos are

---

[1]The source videos of DF-1.0 are from 100 paid actors with informed consents, but its 1,000 target videos are taken from FF++, which are collected from YouTube without explicit consent.

recordings of 100 paid actors, and the 1,000 target videos are adopted from FF++. 1,000 fake videos are synthesized by swapping each of the source identities onto 10 target videos. Instead of using multiple synthesis methods, it adds diversity utilizing augmentation on both real and fake videos with seven perturbation methods. As a result, 50,000 real and 10,000 fake clips are created, respectively. Albeit these databases are considerably larger and more varied compared to their early correspondences, they have not been utilized across Deepfake detection studies as much as FF++ and the DFDC dataset, so their academic validity is yet to be fully established.

3. Korean DeepFake Detection Dataset

3.1. Contributions

3.1.1 Quantity

KoDF incorporates 62,166 unique 90-second-long real clips (62.8 days) and 175,776 unique deepfake clips of 15 seconds or longer (30.5 days). It surpasses the DFDC dataset, the previously largest public deepfake detection database, in terms of both the total duration (38.4 days of source videos and 12.1 days of generated outputs) and the number of clips (48,190 source videos and 104,500 fake videos). In addition, unlike in the DFDC dataset, audio-swapped or augmented clips do not count as synthesized data points in KoDF. To build the fake portion of the dataset, we resort only to the inference output of the six carefully selected synthesis models (details to be discussed in Section 3.3.2) rather than to trivial modifications.

3.1.2 Controlled Subject Distribution

KoDF focuses on a situation in which a person talks to a camera, since it is particularly vulnerable to synthetic modifications, and thus frequently targeted by deepfakes. To maximize diversity of the database, we control the 403 participants' distribution according to age, sex, and recording location as shown in Table 2.

Characteristic		Number	Percentage (%)
		403	100
Age	~19	5	1.24
	20~29	205	50.87
	30~39	106	26.30
	40~49	61	15.14
	50~59	19	4.71
	60~	7	1.74
Sex	Female	205	50.87
	Male	198	49.13
Location	Crowdsourcing	353	87.59
	Studio	50	12.41

Table 2. Subject distribution by age, sex, and recording location.

3.1.3 Quality Assurance and Right Clearance

KoDF is quality-assured via a meticulous inspection process. Every single one of real and deepfake instances is cross-checked by human eyes and ears for likely issues. The details of the process are provided in Section 3.2 and 3.3. The filtering procedure excludes trivial cases such as a complete failure of synthesis that often appears in other deepfake datasets. KoDF thus includes only true threats where the level of realism is so high that a human cannot easily tell if the clip is real or not.

All the real clips of KoDF are solicited from paid participants. We have informed them of the purpose of the database in great detail, emphasizing the possible consequences where their faces may be manipulated and synthesized. All of them agreed to appear in the database and signed a formal agreement. In addition, all the synthesis models employed have been thoroughly examined for potential license issues. If needed, we have asked for the authors' permission to use their models for the database construction.

3.1.4 Forward-Looking

Real-world deepfakes would undergo countless modifications (e.g. compression, resizing, manual editing, etc.) in the process of being generated and shared. Elaborate data augmentation is essential to simulate such transformation [54], and consistency in data makes more controlled augmentation possible. We collect and synthesize full HD videos, equalizing the resolution to 1920×1080. Since it is much easier to downgrade a video in quality than to upgrade, the high-resolution clips of KoDF leave a greater and cleaner room for posterior data augmentation. Expecting various augmentation tricks to be applied, we exempt KoDF from any a priori data augmentation (the adversarial attack on 10% of the data is an addition not a replacement; see section 3.4). This is in contrast to the DFDC dataset and DF-1.0 where various perturbations are an inherent part of data. We intend to leave the choice of optimal augmentation techniques to researchers.

Although *face swap* is the best-known method for deepfake creation, deepfake technologies are not simple equivalents of face-swap neural networks. There are a number of other operating means to fake a person's identity in an image or a video. The most concerning one is *face reenactment*, with which one can manipulate actions and expressions of a person in a video, or even in a still image, with an external video or audio source. [50, 59, 47]. While being actively researched for obvious commercial use cases, it has not received due attention in the field of deepfake detection. We therefore include an extensive amount of reenactment models' output in our database so that future detection models can be more resilient against reenacted deepfakes.

The Korean subjects (and the eight Southeast Asians) in KoDF counterbalance the DFDC dataset, in which the deficiency of East Asian and Southeast Asian is notable (according to the preview of the database, the proportion of East Asian is 9%, and that of Southeast Asian is 3% [24]). The complementary racial composition of KoDF, when put together with other databases, will be critical to build more generalized detection models for real-world applications.

Even in the presence of a working deepfake detection model, the attacking side can take a step further, creating trickier instances to deceive the detector. For example, destructive means to confuse classification models have been devised [29, 28, 27]. Therefore, we add to KoDF adversarial examples to encourage the development of detection models that are robust against such attacks. Refer to Section 3.4 for more details.

3.2. Real Data

Unlike previous deepfake detection databases that are made of found clips [37, 48], the source videos of KoDF are recorded specially to constitute the database. By governing the recording process ourselves, we prevent defective instances and control the distribution with regard to recording environment, emotive content, and speech corpus.

Among the 403 subjects, 353 partake in the *crowdsourcing* task where a subject is asked to film oneself for 150 clips, each of which should last over 90 seconds. The first recording of the 150 clips is an *idle* clip where a subject remains in a natural pose saying nothing. In the half of the remaining 149 recordings, a subject reads an assigned script consisting of 10 sentences. These are *script* clips. The remaining 74 clips are *scenario* clips, in which a subject chooses or makes up a question, and provides his or her responses in a given time. To add more diversity, we introduce minor variations in terms of camera angle, focal length, recording location, background, composition of props, and lighting.

Each clip belongs to one of the three emotive categories: *positive*, *negative*, and *neutral*. The sentences of the script clips and the questions of the scenario clips are formulated as such per recording. The purpose of this task design is to facilitate the recording process for the subjects who are mostly amateurs with no experience in shooting footage of themselves. The aforementioned tactics help vary the subjects' responses while providing them enough materials to continue talking without prolonged pauses.

The sentence corpus for the script clips is comprised of definitions and examples crawled from Standard Korean Language Dictionary [44]. The found sentences are screened if the length is too short or long, or if non-Korean symbols are included. They are subsequently organized by sentence type: *statement*, *question*, and *exclamation*. We adjust the occurrence ratio between the statement type and

the other two types in a script clip to 8:2 for the diversity of expression. The chosen sentences are machine-tagged with respect to their emotion category based on Kunsan National University Korean Sentiment Lexicon [45] by simply accumulating the valence score per token. A script clip contains 10 sentences of the same emotion category; for example, for a positive clip, 10 positive sentences are assigned.

For the scenario clips, 420 questions are collected from amateur writers and proofread. The questions are evaluated for their emotive quality by three annotators and then categorized by the majority rule. In each clip, a subject is asked to choose a question of the corresponding category and answer it during the given time. While a subject is free to make up his or her own question instead of choosing from the question bank, the made-up question needs to meet the emotive category in terms of the content of the question and the expected answer. The purpose of the scenario design is to relax the amateur participants and to widen the range of speaking style, complementing the relatively rigid dynamics and monotonous prosody of the script clips.

The remaining 50 subjects participate in the *studio* task, where the recording environment and the task design differ from those for the crowdsourcing task. They record high-quality video footage at a professional studio with a skilled director against a green screen. A subject carries out eight recording runs in one or two sessions on different days. He or she reads 300 sentences per run, and each of the runs takes approximately 35 minutes. The eight long recordings are later split into 90-second intervals, resulting in 184 clips in total. These clips are equivalent to the script clips in the crowdsourcing task, and we do not include scenario design for the studio task.

The collected real clips are manually inspected for various possible defects: (1) audio-video sync problem, (2) excessive background noise, (3) utterances severely hindered or stuttered, (4) extreme lighting conditions, and (5) face located far outside of the central region. If any of these problems is detected during the checking process, the subject is requested to shoot the corresponding clip again.

3.3. Synthesized Data

We employ six different models to generate deepfake clips. Among them, FaceSwap [9], DeepFaceLab [46], and FSGAN [43] are face swapping models. First Order Motion Model (FOMM) [50] is a *video-driven* face-reenactment model. The remaining two, Audio-driven Talking Face Head Pose (ATFHP) [59] and Wav2Lip [47], are *audio-driven* face-reenactment models. Before and after the actual synthesis, video clips are processed to reduce artifacts and enhance fidelity.

Hereafter, the terms *target* and *source* are used to denote different facial identities in face swapping; the target is the to-be-replaced facial content of the base video, and

the source is the facial content to replace the target. Thus a clip resulting from a face swapping method looks identical to the target video, except for the facial identity which is of the source. On the other hand, the source and the target are one and the same in face-reenactment methods; their goal is to manipulate the pose or expression of the source person while maintaining the rest. The driving audio or video can be taken from different identities.

3.3.1 Preprocessing

All the raw clips are preliminarily processed subject-wise. We employ a facial landmark algorithm 2DFAN [19], by which the facial regions are cropped, aligned, and resized to 512×512 pixels. These regions are computed by an affine transformation of positions centered around eyes and nose. To train face swapping models, 4,000 to 5,000 face frames are selected according to their sharpness and to the diversity of face angles (reenactment models require little or no additional training).

3.3.2 Synthesis Models

The synthesis models of KoDF are a diverse collection of facial manipulation techniques. Although we try to maintain equal distribution amongst the models, since all of the generated clips are validated under manual screening processes, the number of videos per method is not equal. Some methods guarantee a stable level of realism, accounting for a larger chunk while others respond sensitively to lighting and noise, resulting in a number of unusable clips that do not make it into KoDF. Figure 2 shows the example frames generated by the selected methods, and the distribution of the synthesized videos is illustrated in Figure 3.

FaceSwap FaceSwap [9] is an open-source face-swapping software initially developed by a Reddit user */u/deepfakes* and later maintained by its developing community. It bears an encoder-decoder architecture. Two separate decoders handle the source and target faces respectively while sharing one encoder. The three networks are trained concurrently; the encoder learns the non-identity features, while the two decoders capture traits that are more contingent to each of the identities. The model consequently infers face images that maintain the source identity while matching the target's non-identity features. We include this model because of its historical importance as the starting point where the term deepfake originates from.

DeepFaceLab DeepFaceLab [46], currently the most popular means to generate deepfake videos, provides an imperative and easy-to-use pipeline, along with a collection of synthesis models [5]. These models are highly FaceSwap-like, with respect to the convolutional autoencoder architecture as well as the training and inference processes. One of

Figure 2. A selection of synthesized frames. Each row is created using FaceSwap, DeepFaceLab, FSGAN, FOMM, and Wav2Lip from top to bottom.
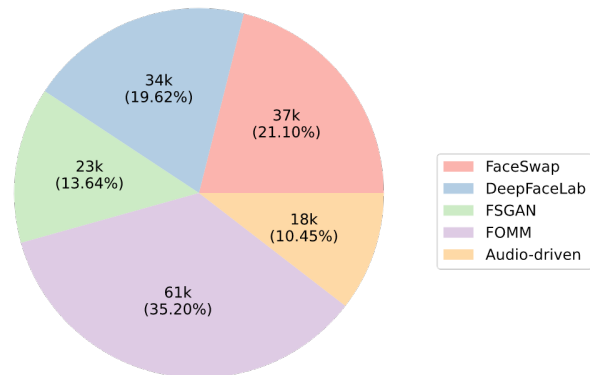

Figure 3. KoDF video distribution by synthesis methods. Audio-driven includes ATFHP and Wav2Lip.

notable improvements from FaceSwap is, however, the intervening network in between the shared encoder and the two decoders. This modification helps capture common non-identity features underlying both the source person and the target person, contributing to robust mapping between the two. In addition, a mixed loss combines structural dissimilarity index with mean squared error, leading to improved fidelity.

FSGAN FSGAN [43] is capable of both face swapping and reenactment. The model first reenacts the source identity according to the target's pose and expression, and segments the facial regions of both faces. Then it inpaints the missing parts of the reenacted face and blends the completed face with the target, creating the final result. During

reenactment, the model selects multiple source frames that correspond most to the target by Delaunay Triangulation, and uses the weighted average of the reenactment results according to barycentric coordinates. This process makes the model subject agnostic so it does not require heavy tuning on every new source. For this paper, we use the face swapping scheme of the official implementation [11]. We adopt the author's recommended swap method for better quality of synthesis[2], which fine-tunes the model for each source-target pair for 800 iterations and utilizes the target mouth area for better teeth quality.

FOMM FOMM [50] is a self supervised network that applies the motion of a driving video sequence to an image where both contain objects of the same category (e.g. faces). It decouples appearance and motion by modeling the movement around the keypoints using affine transformations in a self-supervised manner. It warps the source image based on the motion of the driving video then recovers the warping artifacts by inpainting. We use the official code and the pretrained model [10] on VoxCelab dataset [41]. This model is chosen to represent the face-reenactment strategy and for its real-world applications. For instance, Open Avatarify [12], a popular real-time reenactment tool for video chats, adopts this model.

ATFHP and Wav2Lip We employ two different audio-driven face synthesis models for KoDF: ATFHP [59] and Wav2Lip [22]. The former reflects the active research domain of face synthesis based on 3D morphable model. ATFHP takes audio and video inputs to create an output video, which retains the identity of the input video while synchronizing facial expressions to the audio. This is achieved by creating a list of 3D model parameters from the input audio and render them to synthesized frames. A memory-augmented GAN module then refines the rendered frames into realistic ones with smooth background transition for various face identities. After pretrained with the Lip Reading in the Wild dataset [21], it only requires fine-tuning with a small number of frames to learn personalized talking behavior.

Soon after starting to generate audio-driven face-reenactment instances for KoDF using ATFHP, we experiment with a newly published alternative, Wav2Lip. Unlike similar models that generate a talking face from a driving audio with a GAN-based architecture, Wav2Lip utilizes a pretrained lip-sync discriminator, which helps the model to learn the appropriate lip motion according to the audio. To capture the temporal context of speech, the model uses five consecutive face frames and the respective speech content as input.

While the synthesized results from ATFHP are promising, we switch from ATFHP to Wav2Lip due to the relative

efficiency of the synthesis process of the latter. For every input identity, the pretrained model of ATFHP [1] requires careful fine-tuning to acquire high-fidelity syntheses that meet our quality assurance criterion. On the other hand, that of Wav2Lip [15] can efficiently generate samples of proper quality with respect to unseen facial identities without fine-tuning. As a result, 455(2.5%) and 17,915 (97.5%) clips are synthesized using ATFHP and Wav2Lip, respectively.

3.3.3 Postprocessing

All the methods listed above produce a sequence of image frames matched to the facial region cropped during the preprocessing step. Because most models fail to reconstruct accurate details around the facial boundaries, necessitated is the process of blending the synthesized outcome back into the original frame.

Using the same facial landmark detection [19] from the preprocessing stage, we create a facial mask from the synthesized image frame. The border of the mask region goes under a Gaussian blurring process to reduce the artifacts, and the blurred images are blended into the original video frames of corresponding temporal positions. This postprocessing procedure reduces jitters while preserving details around the facial borders.

3.3.4 Quality Evaluation of KoDF

Once videos are generated from the synthesis models, they go through a manual screening process where they are presented to two raters and subjected to two questions: (1) *Is the clip of high quality?* and (2) *Can the figure in the clip pass as a real human?* We keep only the clips that win both raters' approvals for both of these questions. Each clip is presented once in the size of the horizontal cell phone layout. Raters also examine various technical issues regarding orientation, audio-video synchronization, duration, among others. Clips that fail to pass the screening are simply discarded.

The quality of the synthesized output is evaluated with peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), Fréchet Inception distance (FID) [33], and average keypoint distance (AKD).

$$PSNR(r,g) = 10 \cdot log_{10}(\frac{MAX_I^2}{MSE(r,g)}) \qquad (1)$$

$$SSIM(r,g) = \frac{(2\mu_r\mu_g + c_1)(2\sigma_{rg} + c_2)}{(\mu_r^2 + \mu_g^2 + c_1)(\sigma_r^2 + \sigma_g^2 + c_2)} \qquad (2)$$

$$AKD(r,g) = \frac{1}{P}\sqrt{\sum_{p=1}^{P}(r_p - g_p)^2} \qquad (3)$$

Method	PSNR ↑	SSIM ↑	FID ↓	AKD ↓
FaceSwap	22.10±2.02	0.76±0.05	1.11±0.08	0.21±0.04
DeepFaceLab	21.86±1.82	0.75±0.05	1.12±0.09	0.22±0.04
FSGAN	21.09±2.07	0.79±0.08	1.07±0.09	0.16±0.02
FOMM	26.16±2.96	0.87±0.04	1.00±0.07	0.15±0.02
Audio-driven	24.47±2.39	0.84±0.06	1.09±0.08	0.17±0.02
Total	23.72±3.17	0.81±0.07	1.06±0.09	0.18±0.04

Table 3. Quality evaluation of KoDF for each synthesis method. Audio-driven includes ATFHP and Wav2Lip.

Dataset	PSNR ↑	SSIM ↑	FID ↓	AKD ↓
FF++	23.11±3.22	0.77±0.09	1.11±0.08	0.26±0.08
DFDC	24.54±3.23	0.79±0.08	1.14±0.09	0.25±0.13
DF-1.0	22.15±1.76	0.76±0.06	1.11±0.06	0.19±0.11
KoDF	23.72±3.17	0.81±0.07	1.06±0.09	0.18±0.04

Table 4. Quality comparison of KoDF and other datasets.

For the evaluation, we randomly choose 500 real clips and 500 corresponding synthesized clips. From each of the fake samples, 100 frames are uniformly extracted, and their real matches are taken from the identical temporal positions. Each metric is averaged for these 100 pairs to compute the value for each synthesized clip. Table 3 shows the results by the synthesis methods of KoDF, and Table 4 compares KoDF with the FF++, DFDC, and DF-1.0 datasets.[3]

3.4. Adversarial Attack

In the midst of the burgeoning interest in deepfake detection technologies, researchers are thinking ahead to the next step: adversarial attacks to fool detecting models. Methods that are known to be disruptive for deepfake detection include gradient-based adversarial attacks [27] and intentional noises that hide spatial and spectral artifacts arising from a synthesis model [34]. There is even a toolbox to create such adversarial instances [30], and researchers report the vulnerability of deepfake detection models against black-box attacks [42, 35].

The fast gradient sign method [29] is the means chosen for KoDF to simulate potential malicious attempts to evade detection, for it is the most widely known adversarial method. The process of creating adversarial examples are as follows: we train a preliminary detector model, obtain a sequence of noise frames that confuse the preliminary model, and mix the noise sequence with the corresponding input clip. This strategy is based on the assumption that most detectors are similarly structured as the preliminary detector, which allows the generated adversarial examples to generalize their elusive effects across architectural details of detectors.

We build the preliminary model by topping a pretrained

EfficientNet-B4 [52] base with two fully connected layers. The model is then trained with a portion of KoDF, which is structured as follows: from 10% of randomly selected fake samples, frames between the 150th and 450th positions are extracted at a chance of 0.8%, and from 2% of real clips, one in a hundred frames is stochastically drawn. Approximately 40,000 frames are compiled for each category, and the model is trained for 10 epochs.

To create noised instances, every one in ten clips is randomly selected regardless of its real or fake category and decomposed into a sequence of frames. Each frame is iteratively fed into the trained preliminary model. Obtained is the sign of the generated gradients with respect to the input image x, which is scaled and added to the original frame.

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{4}$$

The scaling factor ϵ for the noise is varied between 0.01, 0.05, and 0.1, and the preliminary model is retrained three times over the course of time, to generate noises of varying nature and intensity. The noised frames are once again put together into a video with the addition of audio from the original source.

4. Detection Evaluation

The ultimate goal of a deepfake detection dataset would be to help develop a general detection model that performs well against a variety of real-world deepfake cases. Most studies on deepfake detection are designed so as to measure how their proposed detection models perform based on a certain deepfake detection dataset. The premise here is that the target deepfake detection dataset is a good approximation of the distribution of real-world deepfake instances.

In the subsequent experiments, we investigate if existing deepfake detection datasets guarantee a sufficient level of generality and how they fare when mixed and tested with out-of-domain data. To this end, we train the winning model of the DFDC competition [14] with combinations of the FF++, DFDC, and KoDF datasets. The multiple variants are then evaluated against unseen test sets, which include the adversarial samples of KoDF (Attack), the GDFD dataset, and DF-1.0.

For fair comparison, the DFDC dataset and KoDF are sampled to match the size of FF++. 1,000 real videos and 4,000 fake videos are randomly chosen from the two datasets. Following the preprocessing method of [14], we use MTCNN [60] to extract a face from every frame of the FF++ and DFDC clips, and frames with unrecognizable faces are ignored. In the case of KoDF, we randomly select 300 frames from each clip to match the total number of face samples per dataset. The extracted facial frames are then divided into training, validation, and test sets for each dataset, with a ratio of 8:1:1. We train the DFDC winning
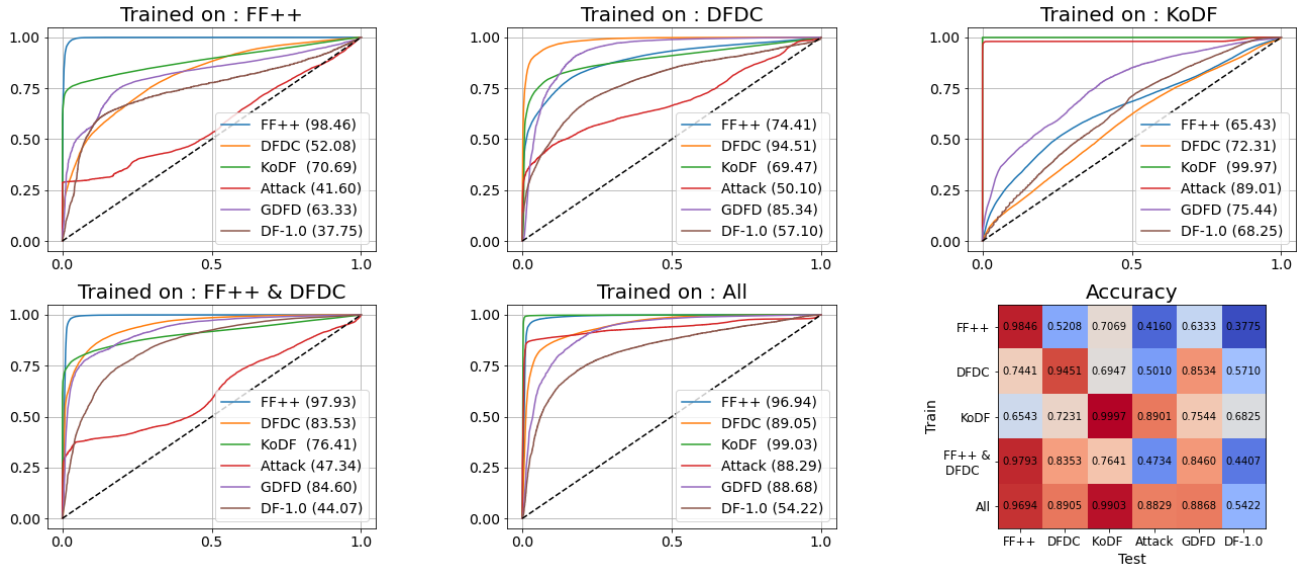
Figure 4. ROC curves of the DFDC winning detection model trained on FF++, the DFDC dataset, KoDF, and their combinations. All training set variants are of equal size. The trained models are evaluated on unseen test sets, including the adversarial samples of KoDF (Attack), the GDFD dataset, and DF-1.0. We also include a corresponding accuracy heat map.

model on different combinations of the three training sets, all of which are of equal size. After 30 epochs of training, the epoch with the minimum loss value on the validation set is chosen, and the trained models are evaluated on each of the six test sets. Figure 4 summarizes the results.

The results show that each of the three deepfake detection datasets, on its own, is not enough to approximate the true distribution of deefakes in the wild, resulting in detection models that are incapable of adapting to different deepfake detection datasets. The issue is relatively severer with KoDF, but it is only natural since the distribution of our dataset is systematically controlled to a greater extent (as detailed in Section 3.1.2 and 3.1.4). The point here is that, when trained on their combinations, the model becomes much more robust against various out-of-domain data. Noteworthy is the performance improvement when the model utilizes all three in comparison to training with only FF++ and the DFDC dataset. This observation supports the complementary utility of KoDF.

From the experimental results, we can deduce that the deepfake detection task is strongly prone to overfitting, much more so than regular image classification tasks where models learn diverse signals recurring naturally (i.e. local patterns and global structures). On the other hand, deepfake detection models focus on artifacts arising during the generation process, which inevitably vary depending on the synthesis methodologies. An ideal deepfake detection dataset should thus incorporate examples of a maximal variety of deepfake methods and a wide range of real videos. No standalone deepfake dataset published so far achieves sufficient generality to meet these conditions on its own, and a practical solution is to utilize multiple datasets adjoined.

5. Conclusions

We present a new large-scale dataset to help researchers develop and evaluate deepfake detection methods. KoDF focuses on Korean subjects to compensate for the Asian underrepresentation of other major deepfake detection databases. It expands the range of employed deepfake methods, regulates the quality of the real clips and the synthesized clips, manages the distribution of subjects according to age, sex, and content, and simulates possible adversarial attacks. While KoDF is an extensive database, our expectation is that it will work even more effectively in the mutual complementation of existing and future deepfake detection databases, including the two milestone datasets FF++ and DFDC. We experimentally demonstrate the benefit of compositing datasets for in-the-wild deepfake detection. We hope KoDF to serve as a stepping stone for future studies in the field of deepfake detection.

6. Acknowledgements

References

[1] Audio-driven talking face head pose official implementation. https://github.com/yiranran/Audio-driven-TalkingFace-HeadPose. Accessed: 2020-07-16.

[2] Avatarify, inc. https://avatarify.ai/. Accessed: 2020-10-17.

[3] Ctrl shift face. https://www.youtube.com/channel/UCKpH0CKltc73e4wh0_pgL3g. Accessed: 2020-09-27.

[4] Deeperforensics challenge 2020. https://competitions.codalab.org/competitions/25228. Accessed: 2020-08-03.

[5] Deepfacelab. https://github.com/iperov/DeepFaceLab. Accessed: 2020-07-16.

[6] Deepfake detection challenge. https://www.kaggle.com/c/deepfake-detection-challenge. Accessed: 2020-05-16.

[7] Deepfake video detecting ai competition (using kodf preview dataset). https://dacon.io/competitions/open/235655/overview/. Accessed: 2020-12-20.

[8] Faceforensics benchmark. http://kaldir.vc.in.tum.de/faceforensics_benchmark/. Accessed: 2020-07-16.

[9] Faceswap. https://github.com/deepfakes/faceswap. Accessed: 2020-06-27.

[10] First order motion model official code. https://github.com/AliaksandrSiarohin/first-order-model. Accessed: 2020-07-16.

[11] Fsgan official implementation. https://github.com/YuvalNirkin/fsgan. Accessed: 2020-07-16.

[12] Open avatarify. https://github.com/alievk/avatarify. Accessed: 2020-09-20.

[13] Reface. https://reface.ai/. Accessed: 2020-10-12.

[14] Selim Seferbekov's dfdc winning model. https://github.com/selimsef/dfdc_deepfake_challenge. Accessed: 2020-09-24.

[15] Wav2lip official code. https://github.com/Rudrabha/Wav2Lip. Accessed: 2020-08-24.

[16] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 38–45, 2019.

[17] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 1205–1207, 2019.

[18] Dan Boneh, Andrew Grotto, Patrick McDaniel, and Nicolas Papernot. Preparing for the age of deepfakes and disinformation. *HAI Policy Brief*, 2020.

[19] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). *IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017.

[20] Matt Burgess. Deepfake porn is now mainstream. and major sites are cashing in. *Wired*, 2020.

[21] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453, 2017.

[22] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. *Computer Vision – ACCV 2016 Workshops*, pages 251–263, 2017.

[23] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[24] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[25] Nick Dufour and Andrew Gully. Contributing data to deepfake detection research. *Google AI Blog*, 2019.

[26] Gereon Fox, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. Videoforensicshq: Detecting high-quality manipulated face videos. *arXiv preprint arXiv:2005.10360*, 2020.

[27] Apurva Gandhi and Shomik Jain. Adversarial perturbations fool deepfake detectors. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[28] Aditya Ganeshan, Vivek B.S., and R. Venkatesh Babu. Fda: Feature disruptive attack. *IEEE International Conference on Computer Vision (ICCV)*, pages 8068–8078, 2019.

[29] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2015.

[30] Dou Goodman, Hao Xin, Wang Yang, Wu Yuesheng, Xiong Junfeng, and Zhang Huan. Advbox: A toolbox to generate adversarial examples that fool neural networks. *arXiv preprint arXiv:2001.05574*, 2020.

[31] Karen Hao. Deepfake porn is ruining women's lives. now the law may finally ban it. *MIT Technology Review*, 2021.

[32] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *Proc. 35th AAAI Conference on Artificial Intelligence Workshops*, 2021.

[33] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6629–6640, 2017.

[34] Yihao Huang, Felix Juefei-Xu, Qing Guo, Xiaofei Xie, Lei Ma, Weikai Miao, Yang Liu, and Geguang Pu. Fakeretouch: Evading deepfakes detection via the guidance of deliberate noise. *arXiv preprint arXiv:2009.09213*, 2020.

[35] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3348–3357, 2021.

[36] Liming Jiang, Wayne Wu, Ren Li, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2895, 2020.

[37] Pavel Korshunov and Sebastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

[38] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. *MM '20: Proceedings of the 28th ACM International Conference on Multimedia*, pages 1864—-1872, 2020.

[39] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deep-fake forensics. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213, 2020.

[40] Iacopo Masi, Aditya Killekar, Royston Mascarenhas, Shenoy Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. *Computer Vision – ECCV 2020)*, pages 667–684, 2020.

[41] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. *Proc. Interspeech 2017*, pages 2616–2620, 2017.

[42] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Ferrer. Adversarial threats to deepfake detection: A practical perspective. *arXiv preprint arXiv:2011.09957*, 2020.

[43] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. *IEEE International Conference on Computer Vision (ICCV)*, pages 7184–7193, 2019.

[44] National Institute of Korean Language. Standard korean language dictionary. `https://stdict.korean.go.kr/main/main.do`. Accessed: 2020-06-25.

[45] SangMin Park, ChulWon Na, MinSeong Cho, DaHee Lee, and ByungWon On. Knu korean sentiment lexicon: Bi-lstm-based method for building a korean sentiment lexicon. *Journal of Intelligence and Information Systems*, 24(4):219–240, 2018.

[46] Ivan Petrov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Um'e, Mr. dpfks, RP Luis, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. Deepfacelab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[47] KR Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. *MM '20: Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[48] Andreas Rössler, Davide Cozzolino, Christian Riess Luisa Verdoliva, Justus Thies, and Matthias Nießner. Face-forensics++: Learning to detect manipulated facial images. *IEEE International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

[49] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 80–87, 2019.

[50] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[51] Catherine Stupp. Fraudsters used ai to mimic ceo's voice in unusual cybercrime case. *The Wall Street Journal*, 2019.

[52] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 97:6105–6114, 2019.

[53] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. Deepfakes evolution: Analysis of facial regions and fake detection performance. *Pattern Recognition. ICPR International Workshops and Challenges*, pages 442–456, 2021.

[54] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.

[55] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, pages 1–13, 2020.

[56] Jonathan Vanian. Why american express is trying technology that makes deepfake videos look real. *Fortune*, 2020.

[57] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot. . . for now. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8692–8701, 2020.

[58] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[59] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yongjin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020.

[60] Kaipeng Zhang, Zhanpeng Zhang, Zifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.