The DeepFake Detection Challenge (DFDC) Dataset

# Brian Dolhansky; Joanna Bitton; Ben Pflaum et al.

## 2020

We show Deepfake detection is extremely difficult and still an unsolved problem, a Deepfake detection model trained only on the DeepFake Detection Challenge can generalize to real "in-the-wild" Deepfake videos, and such a model can be a valuable analysis tool when analyzing potentially Deepfaked videos

# Abstract

Deepfakes are a recent off-the-shelf manipulation technique that allows anyone to swap two identities in a single video. In addition to Deepfakes, a variety of GAN-based face swapping methods have also been published with accompanying code. To counter this emerging threat, we have constructed an extremely large face swap video dataset to enable the training of detection models, and organized the accompanying DeepFake Detection Challenge (DFDC) Kaggle competition. Importantly, all recorded subjects agreed to participate in and have their likenesses modified during the construction of the face-swapped dataset. The DFDC dataset is by far the largest currently and publicly available face swap video dataset, with over 100,000 total clips sourced from 3,426 paid actors, produced with several Deepfake, GAN-based, and non-learned methods. In addition to describing the methods used to construct the dataset, we provide a detailed analysis of the top submissions from the Kaggle contest. We show although Deepfake detection is extremely difficult and still an unsolved problem, a Deepfake detection model trained only on the DFDC can generalize to real "in-the-wild" Deepfake videos, and such a model can be a valuable analysis tool when analyzing potentially Deepfaked videos. Training, validation and testing corpuses can be downloaded from https://ai.facebook.com/datasets/dfdc.

# Study subjects

**3426 subjects**

In order to reflect the potential harm of Deepfaked videos designed to harm a single, possibly non-public person, videos were shot in a variety of natural settings without professional lighting or makeup, (but with high-resolution cameras, as resolution can be easily downgraded). **The source data consisted of: 1. 3,426 subjects in total with an average of 14.4 videos each, with most videos shot in 1080p**. 2. 48,190 total videos that average 68.8s each - a total of 38.4 days of footage

# Findings

Of the total amount of videos, 100,000 clips contained Deepfakes which translates to approximately 83.9% of the dataset being synthetic videos

Similar to the public test set, 50% of the clips included Deepfakes and the other 50% were nonDeepfaked clips

Unlike the public test set, 50% of this dataset includes organic content found on the internet and the other 50% is unseen content from our source video dataset, collected from various sources

Augmentations were applied to approximately 79% of all videos in the final evaluation test set

Augmenters were randomly applied to approximately 70% of the videos and are fairly straightforward transforms

About 30% of all videos contained distractors, some being more adversarial than others

Of all of the scores on the private test set, 60% of submissions had a log loss lower than or equal to 0.69, which is roughly the score if one were to predict a probability of 0.5 for every video

# Scholarcy Highlights

* Swapping faces in photographs has a long history, spanning over one hundred and fifty years [7], as film and digital imagery have a powerful effect on both individuals and societal discourse [15]
* The DeepFake Detection Challenge (DFDC) dataset is by far the largest currentlyand publicly-available face swap video dataset, with over 100,000 total clips sourced from 3,426 paid actors, produced with several Deepfake, GAN-based, and non-learned methods
* Unlike other work in this area, we explicitly do not show the worst examples from datasets other than the DFDC Dataset as a comparison, as (a) it is simple to cherry-pick the worst examples from a distribution of data produced by automatic methods, and (b) the perceptual quality of a moving video cannot be demonstrated with individual still frames
* Face swaps produced with Deepfake Autoencoder (DFAE) methods were of higher quality over a wider range of videos than swaps produced with GAN-like methods, and required much less fine tuning
* Our hypothesis is that GAN-like methods work well in limited settings with even lighting, such as news rooms, interviews, or controlled-capture videos as in [14], but do not work well automatically. This may explain why most public Deepfake videos are produced with DFAE methods
* Similar to the public test set, 50% of the clips included Deepfakes and the other 50% were nonDeepfaked clips
* The second component of this work involved a large public competition, where participants submitted Deepfake detection models trained on the full DFDC Dataset

# Scholarcy Summary

## Introduction

Swapping faces in photographs has a long history, spanning over one hundred and fifty years [7], as film and digital imagery have a powerful effect on both individuals and societal discourse [15].

Producing a Deepfake does not require specialized hardware beyond a consumer-grade GPU, and several off-the-shelf software packages for creating Deepfakes have been released

The combination of these factors has lead to an explosion in their popularity, both in terms of producing parody videos for entertainment, and for use in targeted attacks against individuals or institutions [9].

Detecting Deepfakes at scale necessitates scalable methods, and computer vision or multimodal models are suited to this challenge.

These models require training data, and even though it is possible to create several convincing Deepfakes the cost of producing the hundreds of thousands of Deepfake videos necessary to train these models is often cost prohibitive.

Gathering thousands of submissions and running them against real Deepfake videos that participants never see paints an extremely accurate picture of the true Deepfake detection state of the art

## Previous work

Due to the nature of the pairwise auto-encoder style model used to produce the majority of Deepfake videos, and due to limited availability of source footage, previous datasets contain few videos and fewer subjects.

We propose a third generation of datasets that have more than an order of magnitude larger number of frames and videos than the second generation, and with better quality, and with agreement from individuals appearing in the dataset.

This generation would include the DFDC, as well as the recent DeeperForensics-1.0 (DF-1.0) dataset [14].

The perturbations used in DF-1.0 to expand the original set of 1,000 fake videos only contain basic pixel-level distortions such as color changes and Gaussian noise, and no semantic distractors that are present in real videos

## Source data

Many Deepfake or face swap datasets consist of footage taken in non-natural settings, such as news or briefing rooms.

The subjects in these videos may not agreed to have their faces manipulated

With this understanding, we did not construct our dataset from publicly-available videos.

We commissioned a set of videos to be taken of individuals who agreed to be filmed, to appear in a machine learning dataset, and to have their face images manipulated by machine learning models.

In order to reflect the potential harm of Deepfaked videos designed to harm a single, possibly non-public person, videos were shot in a variety of natural settings without professional lighting or makeup,.

3,426 subjects in total with an average of 14.4 videos each, with most videos shot in 1080p.

2. 48,190 total videos that average 68.8s each - a total of 38.4 days of footage

## Over 25 TB of raw data

The source videos were pre-processed with an internal face tracking and alignment algorithm, and all face frames were cropped, aligned, and resized to 256x256 pixels.

For Deepfake methods, a subsample of 5,000 face frames collected from all videos was used to train models

## Methods

Throughout this section, the terms target and source are used. In general, target refers to the base video in which a face will be swapped; source refers to the source content that is used to extract the identity that will be swapped onto the target video.

The eyes and the mouth are copied from the original videos using blending techniques, and spherical harmonics are used to transfer the illumination

This method works best when both the target and source face expressions are similar, so we used a nearest-neighbors approach on the frame landmarks in order to find the best source/target face pair.

NTH: The NTH [31] model is able to generate realistic talking heads of people in few- and one-shot learning settings

It consists of two distinct training stages: a metalearning stage and a fine-tuning stage.

StyleGAN: The StyleGAN [16] method is modified to produce a face swap between a given fixed identity descriptor onto a video by projecting this descriptor on the latent face space

This process is executed for every frame.

These audio manipulations were not considered ’Deepfakes’ for the competition, we included them in the set to provide more breadth of manipulation type, and they may be of use in further research

## Training

All subjects were split into one of four sets: training, validation, public test, or private test.

The training and validation sets were released publicly, while the public and private test sets were not released, as they were used to rank the final scores of all submissions to the Kaggle contest.

Training all pairs would require almost 1,000 GPUyears, assuming that it takes one day to train a DFAE model on one GPU.

Within a set, subjects were paired with those with similar appearances, as this tended to give better results for models like the DFAE.

A subset of 10 second clips were selected from the output of all models, and the overall distribution of gender and appearance was balanced across all sets and videos

## Post processing

All methods produced a cropped image containing the face at 256x256 resolution.

We blended the face using the mask onto the original frame using Poisson blending.

We did not use Poisson blending over the entire mask, as this would often blend the two identities and create an ”average” face rather than a face that looks like the source subject.

We only blended a small region along the edges of the mask

This was done using a set of morphological operations that extracted the mask border, applying a Gaussian filter to the mask border pixels, and Poisson blending the original and swapped face frames using this transformed mask.

It is important to note that proper face alignment enhanced the quality of all methods.

Faces were aligned by using a triangular set of positions formed by the two eyes and the nose, and computing an affine transform that best aligned a given face with these positions

## Dataset contents

Training set: The training set provided was comprised of 119,154 ten second video clips containing 486 unique subjects.

Validation: The validation set is the public test set used to compute the public leaderboard positions in the Kaggle competition

This dataset consisted of 4,000 ten second video clips, in which 50% (2000 clips) included Deepfakes.

The dataset included one unseen generation method for Deepfakes: StyleGAN.

Similar to the public test set, 50% of the clips included Deepfakes and the other 50% were nonDeepfaked clips.

The half of the final evaluation test set consisting of DFDC videos was assembled using 260 unique subjects from the source video dataset that have not been seen before.

Augmentations were applied to approximately 79% of all videos in the final evaluation test set.

Training, validation and testing corpuses can be downloaded from http://ai.facebook.com (URL to be updated)

## Augmentations

Various augmentations such as geometric transforms or distractors were added to the videos in both the public Kaggle test set as well as the final evaluation test set.

2. Augmenter: applies geometric and color transforms, frame rate changes, etc.

All augmenters were present in both the public and final test sets, except for the grayscale augmenter which was only present in the final evaluation test set.

The simplest distractors overlay random text, shapes, and dots onto each frame of a video and move around frame to frame.

The subset of distractors overlay images onto each frame of a video.

All distractors were present in the final evaluation test set, only the text, shapes, and faces distractors were present in the public Kaggle test set.

## Metrics

In most analyses of a machine learning model’s performance, classification metrics such as log-loss are reported.

The ratio of Deepfaked videos to real videos may be less than one in a million

With such an extreme class imbalance, accuracy is not as relevant as the precision or false positive rate of a model - the number of false positives of even an extremely accurate model will outnumber the true positives, decreasing the utility of the detection model.

Given the large number of true negatives, it is important for an automatic detection model to be precise.

Precision is the most indicative metric for how a detection model will perform over a real distribution of videos.

## Results

As we do not introduce any novel architectures, we describe how well different models and methods perform in practice, and show some of the best and worst examples of each method.

Face swaps produced with DFAE methods were of higher quality over a wider range of videos than swaps produced with GAN-like methods, and required much less fine tuning.

DFAE: The DFAE methods were generally the most flexible and produced the best results out of the methods included in this paper

They were able to handle a variety of lighting conditions and individuals with good temporal coherence, even though inference happened on a frame-byframe basis.

NTH: Of the GAN like methods, this method produced the most consistent quality

It tended to insert similar looking eyes across subjects, regardless of the source ID.

Like other GAN methods, NTH did not produce good results in darker settings.

StyleGAN had trouble matching the illumination in a scene

## Large scale benchmarking

The second component of this work involved a large public competition, where participants submitted Deepfake detection models trained on the full DFDC Dataset.

The public test set was used to rank the public leaderboard while the competition was ongoing.

This set only contained DFDC videos with subjects that never appeared in the dataset.

The ”private” test set included real videos, some of which were Deepfakes, in addition to more DFDC videos that contained even more subjects that hadn’t appeared in any previous set.

The following analysis presents a comprehensive snapshot of the current performance of Deepfake detectors, and in particular, the performance against the private test set gives an idea as to how the best models would perform on a real video distribution

## Meta analysis During the course of the competition, 2,114 teams participated

Teams were allowed to submit two different submissions for final evaluation.

Of all of the scores on the private test set, 60% of submissions had a log loss lower than or equal to 0.69, which is roughly the score if one were to predict a probability of 0.5 for every video.

Good performance on the public test set correlated with good performance on the private test set, as shown in the first image of Figure 6.

All final evaluations were performed on the private test set, using a single V100 GPU.

Submissions had to run over all 10,000 videos in the private test set within 90 hours, but most submissions finished evaluating all videos within 10 total hours, giving a rough average inference time of around 3.6s per-video

## Analysis of submitted models

All scores for all submissions were computed over all videos in the private test set.

The fourth solution, Eighteen Years Old [25], used an ensemble of frame and video models, including EfficientNet, Xception, ResNet [10], and a SlowFast [8] video-based network.

They tailored a score fusion strategy for the DFDC dataset.

The fifth winning solution, The Medics [11], used MTCNN for face detection, as well as an ensemble of 7 models, 3 of which were 3D CNNs, including the I3D model [1]

## Future work

There are three main areas of future work regarding the DFDC Dataset.

We would like to perform a large scale perceptual study of the quality of the videos in the dataset.

We would like to expand the overall size of the dataset.

960 of the roughly 3,500 original identities were included in the dataset, again due to time and computational constraints.

The possibility of releasing the original raw dataset to the research community.

One of the main differences with previous Deepfake datasets is that they do not purport to have agreement from individuals to be included in the datasets.

Releasing all of the roughly 50k 1 minute videos with some additional annotations will help alleviate this problem, and hopefully lead to even higher quality and larger Deepfake datasets in the future

# Builds on previous work

This generation would include the DFDC, as well as the recent DeeperForensics-1.0 (DF-1.0) dataset [14]. **We believe that future face-swapped datasets should seek agreement from individual participants in order to be useful to and ethical for the research community**

# Differs from previous work

Motivated primarily by the fact that many previously-released datasets contained few videos with few subjects and with a limited size and number of methods represented, we wanted to release a dataset with a large number of clips, of varying quality, and with a good representation of current state of the art face swap methods. **Furthermore, as we observed that many publicly-released datasets** [14, 17, 18, 23, 30] did not guarantee that their subjects were willing participants or agreed to have their faces modified, we solicited video data from 3,426 paid actors and actresses speaking in a variety of settings for roughly 15 minutes each

In general, face swaps produced with DFAE methods were of higher quality over a wider range of videos than swaps produced with GAN-like methods, and required much less fine tuning. **Our hypothesis is that GAN-like methods work well in limited settings with even lighting, such as news rooms, interviews, or controlled-capture videos as in** [14], but do not work well automatically (yet)

# Future work

There are three main areas of future work regarding the DFDC Dataset. First, we would like to perform a large scale perceptual study of the quality of the videos in the dataset. Due to time constraints and extenuating circumstances surrounding COVID-19, this portion of the project is delayed, but is ongoing. Second, we would like to expand the overall size of the dataset. Only 960 of the roughly 3,500 original identities were included in the dataset, again due to time and computational constraints. Finally, we are exploring