DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection

Liming Jiang[1] Ren Li[2] Wayne Wu[1,2] Chen Qian[2] Chen Change Loy[1†]
[1]Nanyang Technological University [2]SenseTime Research
liming002@ntu.edu.sg tomo.blade.lee@hotmail.com
wuwenyan@sensetime.com qianchen@sensetime.com ccloy@ntu.edu.sg

Abstract

*We present our on-going effort of constructing a large-scale benchmark for face forgery detection. The first version of this benchmark, DeeperForensics-*1.0*, represents the largest face forgery detection dataset by far, with* 60, 000 *videos constituted by a total of* 17.6 *million frames,* 10 *times larger than existing datasets of the same kind. Extensive real-world perturbations are applied to obtain a more challenging benchmark of larger scale and higher diversity. All source videos in DeeperForensics-*1.0 *are carefully collected, and fake videos are generated by a newly proposed end-to-end face swapping framework. The quality of generated videos outperforms those in existing datasets, validated by user studies. The benchmark features a hidden test set, which contains manipulated videos achieving high deceptive scores in human evaluations. We further contribute a comprehensive study that evaluates five representative detection baselines and make a thorough analysis of different settings.*[1,2]

1. Introduction

Face swapping has become an emerging topic in computer vision and graphics. Indeed, many works [1, 2, 4] on automatic face swapping have been proposed in recent years. These efforts have circumvented the cumbersome and tedious manual face editing processes, hence expediting the advancement in face editing. At the same time, such enabling technology has sparked legitimate concerns, particularly on its potential for being misused and abused. The popularization of "Deepfakes" on the internet has further set off alarm bells among the general public and authorities, in view of the conceivable perilous implications. Accordingly, there is a dire need for countermeasures to be in place promptly, particularly innovations that can effectively detect videos that have been manipulated.

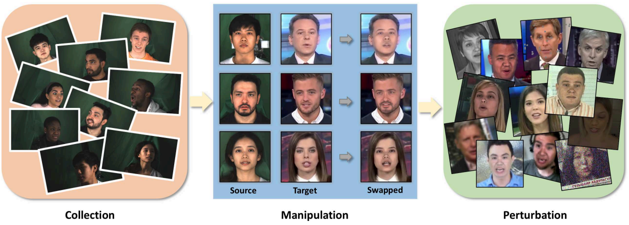

Figure 1: DeeperForensics-1.0 dataset is a new large-scale dataset for *real-world* face forgery detection.

Working towards forgery detection, various groups have contributed datasets (*e.g.*, FaceForensics++ [33], Deep Fake Detection [9] and DFDC [14]) comprising manipulated video footages. The availability of these datasets has undoubtedly provided essential avenues for research into forgery detection. Nonetheless, the aforementioned datasets suffer several drawbacks. Videos in these datasets are either of a small number, of low quality, or overly artificial. Understandably, these datasets are inadequate to train a good model for effective forgery detection in *real-world* scenarios. This is particularly true when current advances in human face editing are able to produce extremely realistic videos, rendering forgery detection a highly challenging task. On another note, we observe high similarity between training and test videos, in terms of their distribution, in certain works [26, 33]. Their actual efficacy in detecting *real-world* face forgery cases, which are much more variable and unpredictable, remains to be further elucidated.

We believe that forgery detection models can only be enhanced when trained with a dataset that is exhaustive enough to encompass as many potential real-world variations as possible. To this end, we propose a large-scale dataset named DeeperForensics-1.0 consisting of 60, 000 videos with a total of 17.6 million frames for real-world face forgery detection. The main steps of our dataset construction are shown in Figure 1. We set forth three yardsticks when constructing this dataset: 1) *Quality.* The dataset shall contain videos more realistic and much closer to the distribution of real-world detection scenarios. (Section 3.1 and 3.2) 2) *Scale.* The dataset shall be made up of a large-scale video sets. (Section 3.3) 3) *Diversity.* There shall be sufficient variations in the video footages (*e.g.*, compression, blurry, transmission errors) to match those that may be en-

[1] GitHub: https://github.com/EndlessSora/DeeperForensics-1.0.
[2] Project page: https://liming-jiang.com/projects/DrF1/DrF1.html.
† Corresponding author.

Table 1: The most relevant datasets compared to our dataset. DeeperForensics-1.0 is an order of magnitude larger in scale than existing datasets *w.r.t.* both real and fake parts. We build a professional indoor environment to better control the important attributes of the collected data. 100 paid actors give consents to the use and manipulation of their faces by signing a formal agreement. We employ seven types of perturbations at five intensity levels, leading to 35 perturbations in total. The video may be subjected to a mixture of more than one perturbation. In contrast to prior works, we also introduce a new end-to-end high-fidelity face swapping method.

Dataset	Total Videos	Ratio (real : fake)	Controlled Capture	Consented Actors	Perturbations (total number)	Perturbations (mixture)	New Method
UADFV [41]	98	1 : 1	×	–	–	×	×
DeepFake-TIMIT [23]	620	only fake	×	–	–	×	×
Celeb-DF [26]	1203	1 : 1.95	×	–	–	×	×
FaceForensics++ [33]	5000	1 : 4	×	–	2	×	×
Deep Fake Detection [9] (joins FaceForensics++)	3431	1 : 8.5	×	28	–	×	×
DFDC Preview Dataset [14]	5214	1 : 3.6	×	66	3	×	×
DeeperForensics-1.0 (Ours)	**60000**	5 : 1	✔	**100**	**35**	✔	✔

countered in the real world (Section 3.3).

The primary challenge in the preparation of this dataset is the lack of good-quality video footages. Specifically, most publicly available videos are shot under an unconstrained environment resulting in large variations, including but not limited to suboptimal illumination, large occlusion of the target faces, and extreme head poses. Importantly, the lack of official informed consents from the video subjects precludes the use of these videos, even for non-commercial purposes. On the other hand, while some videos of manipulated faces are deceptively real, a larger number remains easily distinguishable by human eyes. The latter is often caused by model negligence towards appearance variations or temporal differences, leading to preposterous and incongruous results.

We approach the aforementioned challenge from two perspectives. 1) Collecting fresh face data from 100 individuals with informed consents (Section 3.1). 2) Devising a novel method, DeepFake Variational Auto-Encoder (DF-VAE), to enhance existing videos (Section 3.2). In addition, we introduce diversity into the video footages through deliberate addition of distortions and perturbations, simulating real-world scenarios. We collate the newly collected data and the DF-VAE-modified videos into the DeeperForensics-1.0 dataset, with the aim of further expanding it gradually over time. We benchmark five representative open-source forgery detection methods using our dataset as well as a hidden test set containing manipulated videos that achieve high deceptive ranking in user studies.

We summarize our contributions as follows: 1) We propose a new dataset, DeeperForensics-1.0 that is larger in scale than existing ones, of high quality and rich diversity. To improve its quality, we introduce a carefully designed data collection and a novel framework, DF-VAE, that effectively mitigate obvious fabricated effects of existing manipulated videos. 2) We benchmark results of existing representative forgery detection methods on our dataset, offering insights into the current status and future strategy in face forgery detection.

2. Related Work

Face forgery detection datasets. Building a dataset for forgery detection requires a huge amount of effort on data collection and manipulation. Early forgery detection datasets comprise images captured under highly restrictive conditions, *e.g.*, MICC_F2000 [7], Wild Web dataset [42], Realistic Tampering dataset [24].

Owing to the urgency in video-based face forgery detection, some prominent groups have devoted their efforts to create face forensics video datasets (see Table 1). UADFV [41] contains 98 videos, *i.e.*, 49 real videos from YouTube and 49 fake ones generated by FakeAPP [5]. DeepFake-TIMIT [23] manually selects 16 similar looking pairs of people from VidTIMIT [34] database. For each of the 32 subjects, they generate about 10 videos using low-quality and high-quality versions of faceswap-GAN [4], resulting in a total of 620 fake videos. Celeb-DF [26] includes 408 YouTube videos, mostly of celebrities, from which 795 fake videos are synthesized. FaceForensics++ [33] is the first large-scale face forensic dataset that consists of $4,000$ fake videos manipulated by four methods (*i.e.*, DeepFakes [2], Face2Face [36], FaceSwap [3], NeuralTextures [35])), and $1,000$ real videos from YouTube. Afterwards, Google joins FaceForensics++ and contributes Deep Fake Detection [9] dataset with $3,431$ real and fake videos from 28 actors. Recently, Facebook invites 66 individuals and builds the DFDC preview dataset [14], which includes $5,214$ original and tampered videos with three types of augmentations.

In comparison, we invite 100 paid actors and collect high-resolution (1920×1080) source data with various poses, expressions, and illuminations. 3DMM blendshapes [10] are taken as reference to supplement some exaggerated expressions. We get consents from all the actors for using and manipulating their faces. In contrast to prior works, we also propose a new end-to-end face swapping method (*i.e.*, DF-VAE) and systematically apply seven types of perturbations to the fake videos at five intensity levels. The mixture of distortions to a single video makes our dataset better imitate real-world scenarios. Ultimately, we construct

DeeperForensics-1.0 dataset, which contains 60, 000 videos with 17.6 million frames in total, including 50, 000 original collected videos and 10, 000 manipulated videos.

Face forgery detection benchmarks. A new prominent benchmark, FaceForensics Benchmark [33], for facial manipulation detection has been proposed recently. The benchmark includes six image-level face forgery detection baselines [6, 8, 12, 13, 15, 32]. Although FaceForensics Benchmark adds distortions to the videos by converting them into different compression rates, a deeper exploration of more perturbation types and their mixture is missing. Celeb-DF [26] also provides a face forgery detection benchmark including seven methods [6, 12, 25, 28, 30, 41, 43] trained and tested on different datasets. In aforementioned benchmarks, the test set usually shares a similar distribution with the training set. Such an assumption inherently introduces biases and renders these methods impractical for face forgery detection in real-world settings with much more diverse and unknown fake videos.

In our benchmark, we introduce a challenging hidden test set with manipulated videos achieving high deceptive scores in user studies, to better simulate *real-world* distribution. Various perturbations are analyzed to make our benchmark more comprehensive. In addition, we mainly exploit *video-level* forgery detection baselines [11, 16, 17, 37, 38]. Temporal information – a significant cue for video forgery detection besides single-frame quality – has been considered. We will elaborate our benchmark in Section 4.

3. A New Large-Scale Face Forensics Dataset

The main contribution of this paper is a new large-scale dataset for real-world face forgery detection, DeeperForensics-1.0, which provides an alternative to existing databases. To construct a dataset more suitable for real-world face forgery detection, we design this dataset with careful consideration of *quality*, *scale*, and *diversity*. In Section 3.1 and 3.2, we will discuss the details of data collection and methodology (*i.e.*, DF-VAE) to improve *quality*. In Section 3.3, we will show how to ensure large *scale* and high *diversity* of DeeperForensics-1.0.

3.1. Data Collection

Source data is the first factor that highly affects *quality*. Taking results in Figure 2 as an example, the source data collection increases the robustness of our face swapping method to extreme poses, since videos on the internet usually have limited head pose variations.

We refer to the identity in the driving video as the “target” face and the identity of the face that is swapped onto the driving video as the “source” face. Different from previous works, we find that the source faces play a much more critical role than the target faces in building a high-quality dataset. Specifically, the expressions, poses, and lighting conditions of source faces should be much richer in order to perform robust face swapping. Hence, our data collection mainly focuses on source face videos. Figure 3 shows the diversity in different attributes of our data collection.

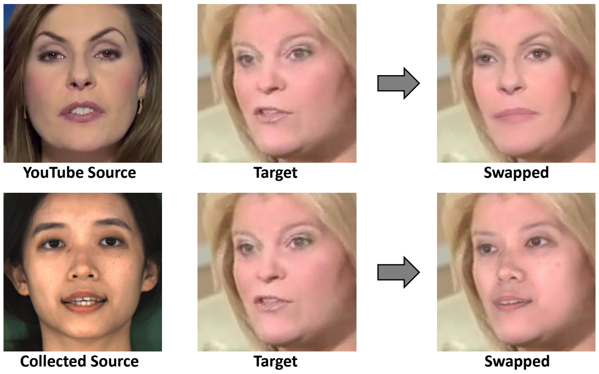

Figure 2: Comparison of using only YouTube video and the collected video as source data, with the same method and setting.

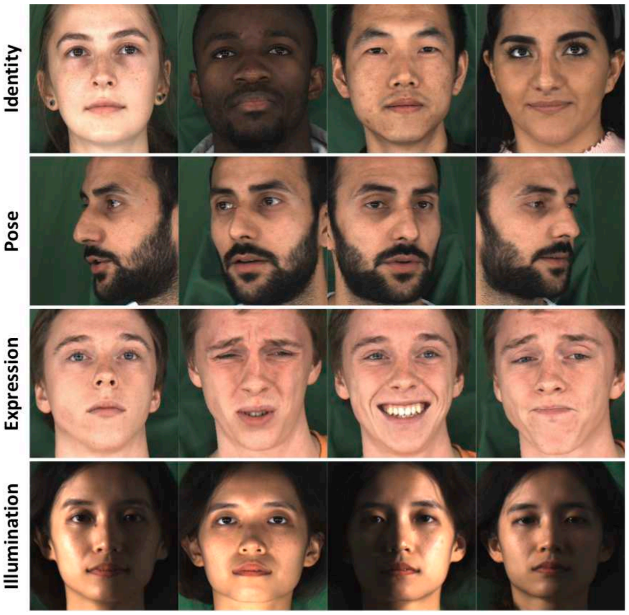

Figure 3: Diversity in identities, poses, expressions, and illuminations in our collected source data.

We invite 100 paid actors to record the source videos. Similar to [9, 14], we obtain consents from all the actors for using and manipulating their faces to avoid the portrait right issues. The participants are carefully selected to ensure variability in genders, ages, skin colors, and nationalities. We maintain a roughly equal proportion *w.r.t.* each of the attributes above. In particular, we invite 53 males and 47 females from 26 countries. Their ages range from 20 to 45 years old to match the most common age group appearing on real-world videos. The actors have four typical skin tones: *white*, *black*, *yellow*, *brown*, with ratio 1:1:1:1. All faces are clean without glasses or decorations.

Different from previous data collection in the wild (see Table 1), we build a professional *indoor* environment for a more controllable data collection. We only use the facial

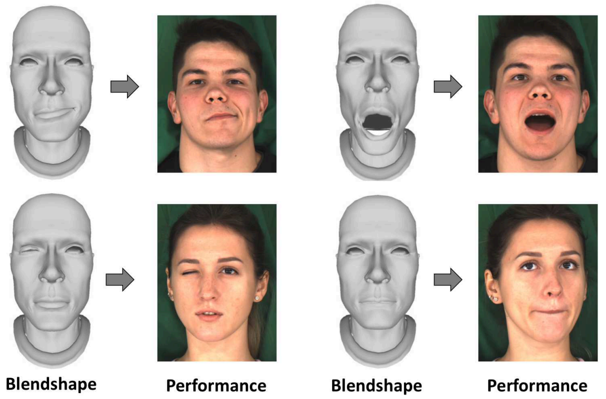

Figure 4: Examples of 3DMM blendshapes in our data collection.

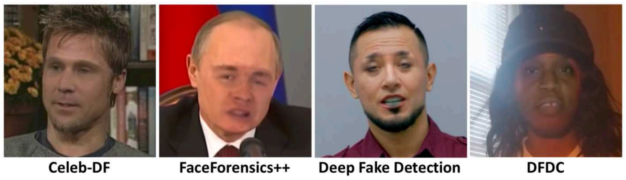

Figure 5: Examples of style mismatch problems in prominent face forensics datasets.

regions (detected and cropped by LAB [40]) of the source data, so we can neglect the background. We set seven HD cameras from different angles: front, left, left-front, right, right-front, oblique-above, oblique-below. The resolution of our recorded videos is high (1920 × 1080). We train the actors in advance to keep the collection process smooth. We request the actors to turn their heads and speak naturally with eight expressions: neutral, angry, happy, sad, surprise, contempt, disgust, fear. The head poses range from -90° to $+90^\circ$. Furthermore, the actors are asked to perform 53 expressions defined in 3DMM blendshapes [10] (see Figure 4) to supplement some extremely exaggerated expressions. When performing 3DMM blendshapes, the actors also speak naturally to avoid excessive frames that show a closed mouth. In addition to expressions and poses, we systematically set nine lighting conditions from various directions: uniform, left, top-left, bottom-left, right, top-right, bottom-right, top, bottom. The actors are only asked to turn their heads under uniform illumination, so the lighting remains unchanged on specific facial regions to avoid many duplicated data samples recorded by the cameras set at different angles. In the end, our collected data contain over $50,000$ videos with a total of 12.6 million frames – an order of magnitude more than existing datasets.

3.2. DeepFake Variational Auto-Encoder

To tackle low visual *quality* problems of previous works, we consider three key requirements in formulating a high-fidelity face swapping method: 1) It should be general and scalable for us to generate large number of videos with high quality. 2) The problem of face style mismatch caused by appearance variations need to be addressed. Some failure cases of existing methods are shown in Figure 5. 3) Temporal continuity of generated videos should be considered.

Based on the aforementioned requirements, we propose DeepFake Variational Auto-Encoder (DF-VAE), a novel learning-based face swapping framework. DF-VAE consists of three main parts, namely a structure extraction module, a disentangled module, and a fusion module. We will give a brief and intuitive introduction of the DF-VAE framework below. Please refer to the *supplementary material* for detailed derivations and results.

Disentanglement of structure and appearance. The first step of our method is face reenactment – animating the source face with similar expression as the target face, without any paired data. Face swapping is considered as a subsequent step of face reenactment that performs blending between the reenacted face and the target background. For robust and scalable face reenactment, we should cleanly disentangle structure (*i.e.*, expression and pose) and appearance representation (*i.e.*, texture, skin color, *etc.*) of a face. This disentanglement is rather difficult because structure and appearance representation are far from being independent.

The blue arrows in Figure 6 show the reconstruction procedure of the source face x_t. Instead of feeding a single source face x_t, we sample another source face x' to construct unpaired data in the source domain. To make the structure representation more evident, we use the stacked hourglass networks [29] to extract landmarks of x_t in the structure extraction module and get the heatmap $\hat{x}_t$. Then we feed the heatmap $\hat{x}_t$ to the Structure Encoder E_α, and x' to the Appearance Encoder E_β. We concatenate the latent representations (small cubes in red and green) and feed it to the Decoder D_γ. Finally, we get the reconstructed face $\tilde{x}_t$, *i.e.*, marginal log-likelihood of x_t. In the target face domain, the reconstruction procedure is the same, as shown by orange arrows.

During training, the network learns structure and appearance information in both the source and the target domains. Exploiting the reparameterization trick [22], the non-differentiable operation of sampling can be made differentiable by an auxiliary variable with independent marginal. The approximate posterior is estimated by the separated encoders E_α and E_β in an end-to-end training process by standard gradient descent. It is noteworthy that even if both y_t and x' belong to arbitrary identities, our effective disentangled module is capable of learning meaningful structure and appearance information of each identity.

During inference, we concatenate the appearance prior of x' and the structure prior of y_t (small cubes in red and orange) in the latent space. The reconstructed face d_t shares the same structure with y_t and keeps the appearance of x'. Our framework allows concatenations of structure and appearance latent codes extracted from arbitrary identities in inference and permits *many-to-many face reenactment*.

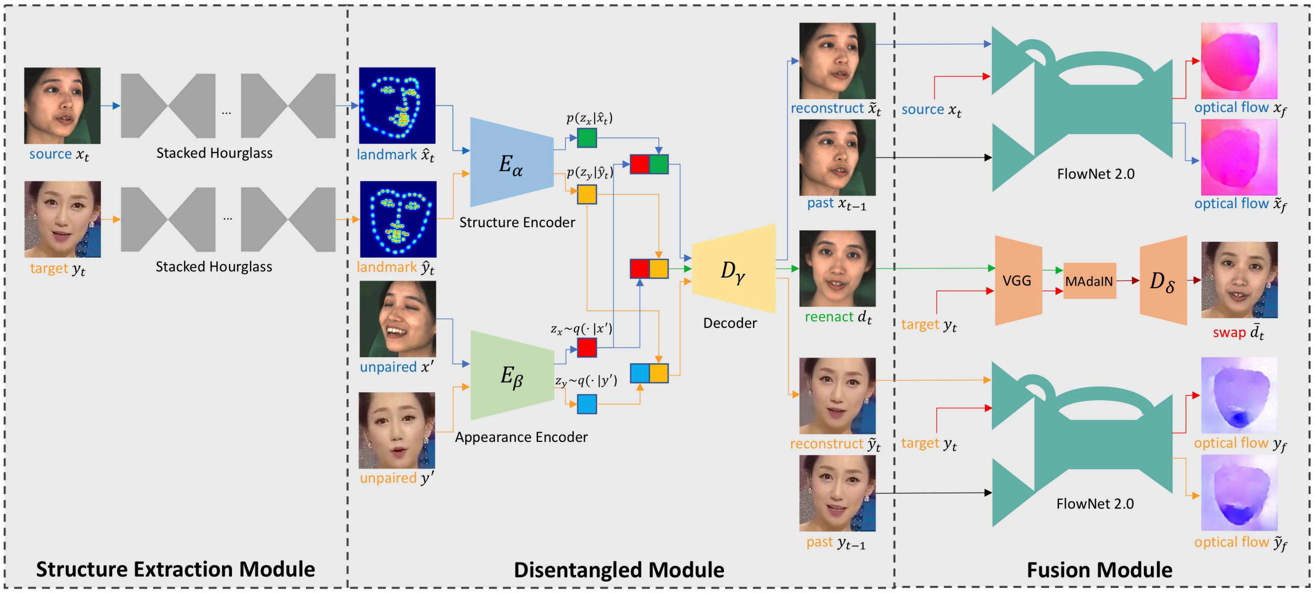

Figure 6: The main framework of DeepFake Variational Auto-Encoder. In training, we reconstruct the source and target faces in blue and orange arrows, respectively, by extracting landmarks and constructing an unpaired sample as the condition. Optical flow differences are minimized after reconstruction to improve temporal continuity. In inference, we swap the latent codes and get the reenacted face in green arrows. Subsequent MAdaIN module fuses the reenacted face and the original background resulting in the swapped face.

In summary, DF-VAE is a new conditional variational auto-encoder [21] with high robustness and scalability. It conditions on two posteriors in different domains. In the disentangled module, the separated design of two encoders E_α and E_β, the explicit structure heatmap, and the unpaired data construction jointly force E_α to learn structure information and E_β to learn appearance information.

Style matching and fusion. To fix the obvious style mismatch problems as shown in Figure 5, we introduce a masked adaptive instance normalization (MAdaIN) module. We place a typical AdaIN [18] network after the reenacted face d_t. In the face swapping scenario, we only need to adjust the style of the face area and use the original background. Therefore, we use a mask m_t to guide AdaIN [18] network to focus on style matching of the face area. To avoid boundary artifacts, we apply Gaussian Blur to m_t and get the blurred mask m_t^b.

In our face swapping context, d_t is the content input of MAdaIN, y_t is the style input. MAdaIN adaptively computes the affine parameters from face area of the style input:

$$\mathrm{MAdaIN}\left(\mathrm{c}, \mathrm{s}\right) = \sigma\left(\mathrm{s}\right)\left(\frac{\mathrm{c} - \mu\left(\mathrm{c}\right)}{\sigma\left(\mathrm{c}\right)}\right) + \mu\left(\mathrm{s}\right), \quad (1)$$

where $c = m_t^b \cdot d_t$, $s = m_t^b \cdot y_t$. With the very low-cost MAdaIN module, we reconstruct d_t again by Decoder D_δ. The blurred mask m_t^b is used again to fuse the reconstructed image with the background of y_t. At last, we get the swapped face $\overline{d}_t$. Figure 7 shows the effectiveness of MAdaIN module for style matching and fusion.

The MAdaIN module is jointly trained with the disentangled module in an end-to-end manner. Thus, by a *single* model, DF-VAE can perform *many-to-many face swapping* with obvious reduction of style mismatch and facial boundary artifacts (see *supplementary material*). Even if there are multiple identities in both the source domain and the target domain, the quality of face swapping does not degrade.

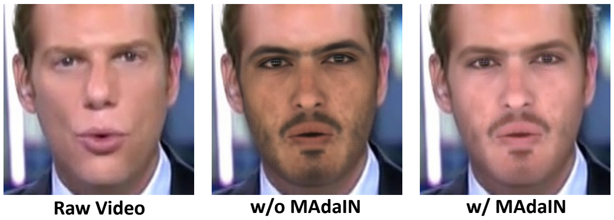

Figure 7: Comparison of the swapped face styles without or with MAdaIN module.

Temporal consistency constraint. Temporal discontinuity of fake videos leads to obvious flickering of the face area, making them very easy to be spotted by forgery detection methods and human eyes. To improve temporal continuity, we let the disentangled module to learn temporal information of both the source face and the target face.

For simplification, we make a Markov assumption that the generation of the frame at time t sequentially depends on its previous P frames $\mathbf{x}_{(t-p):(t-1)}$. We set $P = 1$ to balance quality improvement and training time.

To build the relationship between the current frame and previous ones, we further make an assumption that the optical flows should remain unchanged after reconstruction. We use FlowNet 2.0 [19] to estimate the optical flow $\tilde{x}_f$ *w.r.t.* $\tilde{x}_t$ and x_{t-1}, x_f *w.r.t.* x_t and x_{t-1}. Since face swapping is sensitive to minor facial details that can be greatly affected

Table 2: Seven types of distortions in DeeperForensics-1.0.

No.	Distortion Type
1	Change of color saturation
2	Local block-wise distortion
3	Change of color contrast
4	Gaussian blur
5	White Gaussian noise in color components
6	JPEG compression
7	Change of video constant rate factor

by flow estimation, we do not warp x_{t-1} by the estimated flow like [39]. Instead, we minimize the difference between $\tilde{x}_f$ and x_f to improve temporal continuity while keeping facial detail generation stable. To this end, we propose a new temporal consistency constraint, which can be written as:

$$L_{temporal} = \frac{1}{CHW}\|\tilde{x}_f - x_f\|_1, \tag{2}$$

where $C = 2$ for a common form of optical flow.

We only discuss the temporal continuity *w.r.t.* the source face because the case of the target face is the same. If multiple identities exist in one domain, temporal information of all these identities can be learned in an end-to-end manner.

3.3. Scale and Diversity

Our extensive data collection and the proposed DF-VAE method are designed to improve the *quality* of manipulated videos in DeeperForensics-1.0 dataset. In this section, we will mainly discuss the *scale* and *diversity* aspects. We provide $10,000$ manipulated videos with 5 million frames. It is an order of magnitude more than the previous datasets. We take $1,000$ refined YouTube videos collected by FaceForensics++ [33] as the target videos. Each face of our collected 100 identities is swapped onto 10 target videos, thus $1,000$ raw manipulated videos are generated directly by DF-VAE in an end-to-end process. Thanks to the scalability and multimodality of DF-VAE, the time overhead of model training and data generation is reduced to $1/5$ compared to the common Deepfakes methods, with no degradation in quality. Thus, a larger-scale dataset construction is possible.

To ensure *diversity*, we apply various perturbations to better simulate videos in real scenes. Specifically, as shown in Table 2, seven types of distortions defined in Image Quality Assessment (IQA) [27, 31] are included. Each of these distortions is divided into five intensity levels. We apply random-type distortions to the $1,000$ raw manipulated videos at five different intensity levels, producing a total of $5,000$ manipulated videos. Besides, an additional of $1,000$ robust manipulated videos are generated by adding random-type, random-level distortions to the $1,000$ raw manipulated videos. Moreover, in contrast to all the previous datasets, each sample of another $3,000$ manipulated videos in DeeperForensics-1.0 is subjected to a mixture of more than one distortion. The variability of perturbations improves the *diversity* of DeeperForensics-1.0 to better imitate the data distribution of real-world scenarios. The 10,000 manipulated videos, together with the 50,000 *high-quality* source videos, form the proposed DeeperForensics-1.0.

Table 3: The percentage of user study ratings for UADFV, DeepFake-TIMIT, Celeb-DF, FaceForensics++, Deep Fake Detection, DFDC, and DeeperForensics-1.0 dataset. A higher score means the users think the videos are more realistic.

Dataset	1	2	3	4	5	"real"
UADFV [41]	29.2	36.0	20.7	8.9	5.2	14.1%
DeepFake-TIMIT [23]	31.4	31.4	24.8	9.6	2.7	12.3%
Celeb-DF [26]	5.6	14.8	18.6	24.2	36.9	61.0%
FaceForensics++ [33]	46.8	31.4	13.4	4.4	4.0	8.4%
Deep Fake Detection [9]	26.0	28.0	24.1	11.5	10.3	21.9%
DFDC [14]	25.4	29.7	22.0	11.9	11.1	23.0%
DeeperForensics-1.0 (Ours)	4.3	8.9	22.6	29.8	34.3	**64.1%**

3.4. User Study

To examine the quality of DeeperForensics-1.0 dataset, we engage 100 professional participants, most of whom specialize in computer vision research. We believe these participants are qualified and well-trained in assessing realness of tempered videos. The user study is conducted on DeeperForensics-1.0 and six former datasets, *i.e.*, UADFV [41], DeepFake-TIMIT [23], Celeb-DF [26], FaceForensics++ [33], Deep Fake Detection [9], DFDC [14]. We randomly select 30 video clips from each of these datasets and prepare a platform for the participants to evaluate their realness. Similar to the user study of [20], the participants are asked to provide their feedbacks to the statement "The video clip looks real." and give scores at five levels (1-clearly disagree, 2-weakly disagree, 3-borderline, 4-weakly agree, 5-clearly agree. We assume that users who give a score of 4 or 5 think the video is "real"). The user study results are presented in Table 3. The quality of our dataset is appreciated by most of the participants. Compared to the previous datasets, DeeperForensics-1.0 achieves the highest realism rating. Although Celeb-DF [26] also gets very high realness scores, the scale of our dataset is much larger.

4. Video Forgery Detection Benchmark

Dataset split. We exploit $1,000$ raw manipulated videos in Section 3.3 and $1,000$ YouTube videos from FaceForensics++ [33] as our *standard* set. The videos are split into training, validation, and test set with a ratio of $7:1:2$. The identities of the swapped faces may be duplicated because faces of 100 invited actors are swapped onto $1,000$ driving videos. To avoid data leak, we randomly choose unrepeated 70, 10, and 20 identities, and group all the videos according to the identities. Similar to [33], the test and training sets share a close distribution in our *standard* set. Other experiments in our benchmark are conducted on different variants of the standard set. These variants share the same $1,000$ driving videos with the standard set. We will detail them in Section 4.2. For a fair comparison, all the experiments are conducted in the same split setting.

Hidden test set. For real-world scenarios, some experiments in prior works [26, 33] may not perform a convincing evaluation due to the huge biases caused by a close distribution between the training and test sets. The aforementioned standard set has the same setting. As a result, strong detection baselines obtain very high accuracy on the standard test set as demonstrated in Section 4.2. However, the ultimate goal of face forensics datasets is to help detect forgeries in real scenes. Despite the high accuracy on the standard test set, the models may easily fail in real-world scenarios.

We argue that the test set of *real-world* face forgery detection *should not* share a close distribution with the training set. We need a test set that better simulates real-world settings. We call it "hidden" test set. The hidden test set should satisfy three factors: 1) *Multiple sources.* Fake videos in-the-wild should be manipulated by different unknown methods. 2) *High quality.* Threatening fake videos should have high quality to fool human eyes. 3) *Diverse distortions.* Different perturbations should be taken into consideration.

Thus, in our initial benchmark, we introduce a challenging hidden test set with 400 carefully selected videos. First, we collect fake videos generated by several unknown face swapping methods to ensure multiple sources. Then, we obscure all selected videos multiple times with diverse hidden distortions that are commonly seen in real scenes. Finally, we only select videos that can fool at least 50 out of 100 human observers in a user study. The ground truth labels are hidden and are used on our host server to evaluate the accuracy of detection models. Besides, the hidden test set will be enlarged constantly to get future versions along with development of Deepfakes technology. Fake videos manipulated by future face swapping methods will be included as long as they can pass the human test supported by us.

4.1. Baselines

Existing studies [26, 33] primarily provide image-level face forgery detection benchmark. However, fake videos in-the-wild are much more menacing than manipulated images. We propose to conduct evaluation mainly based on video classification methods for two reasons. First, image-level face forgery detection methods do not consider any temporal information – an important cue for video-based tasks. Second, image-level methods have been widely studied. We only choose one image-level method, XceptionNet [12], which achieves the best performance in [33], as one part of our benchmark for reference. The other four video-based baselines are C3D [37], TSN [38], I3D [11], and ResNet+LSTM [16, 17], all of which have achieved promising results in video classification tasks. Details of all the baselines can be found in our *supplementary material*.

4.2. Results and Analysis

Owing to the goal of detecting fakes in real-world scenarios, we mainly explore how real-world distortions affect the model performance. Accuracies of face forgery detection on the standard test set and the introduced hidden test set are evaluated under various settings.

Table 4: The binary detection accuracy of the baselines on the hidden test set when trained on four manipulated methods in FaceForensics++ (FF++): DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT), and on DeeperForensics-1.0 standard training set without distortions.

Train	FF++ DF	FF++ F2F	FF++ FS	FF++ NT	**DeeperForensics-1.0**
Test (acc)	hidden	hidden	hidden	hidden	**hidden**
C3D [37]	57.50	57.75	52.13	58.25	**74.75**
TSN [38]	57.63	57.25	53.50	57.38	**77.00**
I3D [11]	56.63	58.38	54.63	63.63	**79.25**
ResNet+LSTM [16, 17]	57.38	56.13	54.88	59.50	**78.25**
XceptionNet [12]	57.38	58.75	54.75	57.38	**77.00**

Evaluation of effectiveness of DeeperForensics-1.0. For a fair comparison, we evaluate DeeperForensics-1.0 and the state-of-the-art FaceForensics++ [33] dataset because they use the same driving videos. In this setting, we use $1,000$ raw manipulated videos without distortions in the standard set of DeeperForensics-1.0. For FaceForensics++, the same split is applied to its *four* subsets. All the models are tested on the hidden test set (see Table 4).

The baselines trained on the standard training set of DeeperForensics-1.0 achieve much better performance on the hidden test set than all the *four* subsets of FaceForensics++. This proves the higher *quality* of DeeperForensics-1.0 over prior works, making it more useful for real-world face forgery detection. In Table 4, I3D [11] obtains the best performance on the hidden test set when trained on the standard training set. We conjecture that the temporal discontinuity of fake videos leads to higher accuracy by this video-level forgery detection method.

Evaluation of dataset perturbations. We study the effect of perturbations towards the forgery detection model performance. In contrast to prior work [33], we try to evaluate the baseline accuracies when applying different distortions to the training and the test sets, in order to explore the function of perturbations in face forensics dataset.

In this setting, we conduct all the experiments on DeeperForensics-1.0 dataset with high diversity of perturbations. We use $1,000$ manipulated videos in the standard set (std), $1,000$ manipulated videos with single-level (level-5), random-type distortions (std/sing), $1,000$ manipulated videos with random-level, random-type distortions (std/rand). The data split is the same as that of the standard set with a ratio of $7:1:2$.

In Column 2 of Table 5, we find the accuracy is nearly 100% when the models are trained and tested on the standard set. This is reasonable because the strong baselines perform very well in a clean dataset with the same distribution. In Columns 3 and 4, the accuracy decrease compared to Column 2, when we choose std/sing and std/rand as the test set. Most of the video-level methods except C3D [37] are more robust to perturbations on test set than XceptionNet [12]. This setting is very common because different distributions of the training and the test sets lead to decrease

Table 5: The binary detection accuracy of the baselines when trained and tested on DeeperForensics-1.0 dataset with different distortion perturbations. We analyze different training and testing settings on the standard set without distortions (std), the standard set with single-level distortions (std/sing), and the standard set with random-level distortions (std/rand).

Train Test (acc)	**std** **std**	std std/sing	std std/rand	std/sing std/sing	std/rand std/rand	std/sing std/rand	std/rand std/ sing
C3D [37]	**98.50**	87.63	92.38	95.38	96.63	96.75	94.00
TSN [38]	**99.25**	91.50	95.00	98.25	98.88	98.12	99.12
I3D [11]	**100.00**	90.75	96.88	99.50	99.63	99.63	98.00
ResNet+LSTM [16, 17]	**100.00**	90.63	97.13	100.00	98.63	100.00	97.25
XceptionNet [12]	**100.00**	88.38	94.75	99.63	99.63	99.75	99.00

in model accuracies. Hence, the lack of perturbations in the face forensics dataset cutbacks the model performance for real-world face forgery detection with even more complex data distribution.

When we apply corresponding distortions to the training and test sets, the accuracy will increase (Column 5 and 6 in Table 5) compared to Column 3 and 4. However, this setting is impractical because the distributions of the training and test sets are still the same. We should augment the test set to better simulate the real-world distribution. Thus, some evaluation settings in previous works [26, 33] are unreasonable. If we swap the training set and the test set of std/sing and std/rand to further randomize the condition, results shown in Column 7 and 8 indicate that the accuracy remains high. This evaluation setting shows the possibility that with the same generation method, exerting appropriate distortions to the training set can make face forgery detection models more robust to real-world perturbations.

Evaluation of variants of training set for real-world face forgery detection. We have conducted several experiments for evaluations of possible perturbations. Nevertheless, the case is more complex in real scenes because no information about the fake videos is available. The video may be subjected to more than one type and diverse levels of distortions. In addition to distortions, the method manipulating the faces is unknown.

From the evaluation of perturbations, we find the possibility of augmenting the training set to improve detection model performance. Thus, we further evaluate baseline performance on the hidden test set by devising some variants of the training set. We perform experiments on DeeperForensics-1.0. In this setting, other than std, std/sing, and std/rand, we use additional $1,000$ manipulated videos, each of which is subjected to a mixture of three random-level, random-type distortions (std/mix). We combine std with std/sing, std/rand, and std/mix, respectively, yielding three new training sets (with the same data split as the former settings).

Column 2 in Table 6 shows the low accuracy when the models trained on std and tested on the hidden test set (same as Column 6 in Table 4). Columns 3 and 4 indicate that the accuracy of all the baseline models increase when trained on std+std/sing and std+std/rand. The accuracy of I3D [11] and ResNet+LSTM [16, 17], are over 80% in some cases.

Table 6: The binary detection accuracy of the baselines on the hidden test set when trained on DeeperForensics-1.0 dataset with the standard set without distortions (std), combination of std and the standard set with single-level distortions (std+std/sing), combination of std and the standard set with random-level distortions (std+std/rand), combination of std and the standard set with the mixed distortions(std+std/mix).

Train Test (acc)	std hidden	std+std/sing hidden	std+std/rand hidden	**std+std/mix** **hidden**
C3D [37]	74.75	78.25	78.13	**78.88**
TSN [38]	77.00	78.75	79.50	**79.50**
I3D [11]	79.25	80.13	80.13	**80.13**
ResNet+LSTM [16, 17]	78.25	80.25	79.50	**80.25**
XceptionNet [12]	77.00	79.75	79.75	**79.88**

In a more complex setting, when the models are trained on std+std/mix, Column 5 shows the accuracy of all the detection baselines further increase.

The results suggest that designing suitable training set variants has the potential to help increase face forgery detection accuracy, and applying various distortions to ensure the *diversity* of DeeperForensics-1.0 is necessary. In addition, compared to image-level method, video-level face forgery detection methods have more potential capabilities to crack real-world fake videos as shown in Table 6. Although the accuracy on the challenging hidden test set is still not very high, we provide two initial directions for future real-world face forgery detection research: 1) Improving the source data collection and generation method to ensure the *quality* of the training set; 2) Augmenting the training set by various distortions to ensure its *diversity*. We welcome researchers to make our benchmark more comprehensive.

5. Discussion

In this work, we propose a new large-scale dataset named DeeperForensics-1.0 to facilitate the research of face forgery detection towards *real-world* scenarios. We make several efforts to ensure *good quality*, *large scale*, and *high diversity* of this dataset. Based on the dataset, we further benchmark existing representative forgery detection methods, offering insights into the current status and future strategy in face forgery detection. Several topics can be considered as future works. 1) We will continue to collect more source and target videos to further expand DeeperForensics. 2) We plan to invite interested researchers for contributing their video falsification methods to enlarge our hidden test set, as long as the fakes can pass the human test supported by us. 3) A better evaluation metric for face forgery detection methods is also an interesting research topic.

Acknowledgments. This work is supported by the SenseTime-NTU Collaboration Project, Singapore MOE AcRF Tier 1 (2018-T1-002-056), NTU SUG, and NTU NAP. We gratefully acknowledge the exceptional help from Hao Zhu and Keqiang Sun for their contribution on source data collection and coordination.

References

[1] Deepfacelab. https://github.com/iperov/DeepFaceLab/. Accessed: 2019-08-20. 1

[2] Deepfakes. https://github.com/deepfakes/faceswap/. Accessed: 2019-08-16. 1, 2

[3] Faceswap. https://github.com/MarekKowalski/FaceSwap/. Accessed: 2019-08-18. 2

[4] faceswap-gan. https://github.com/shaoanlu/faceswap-GAN/. Accessed: 2019-08-16. 1, 2

[5] Fakeapp. https://www.fakeapp.com/. Accessed: 2019-07-25. 2

[6] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *WIFS*, 2018. 3

[7] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra. A sift-based forensic method for copy–move attack detection and transformation recovery. *TIFS*, 6:1099–1110, 2011. 2

[8] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *IH & MMSEC*, 2016. 3

[9] Google AI Blog. Contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. Accessed: 2019-09-25. 1, 2, 3, 6

[10] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *TVCG*, 20:413–425, 2013. 2, 4

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 3, 7, 8

[12] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 3, 7, 8

[13] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *IH & MMSEC*, 2017. 3

[14] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint*, arXiv:1910.08854, 2019. 1, 2, 3, 6

[15] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *TIFS*, 7:868–882, 2012. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 7, 8

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–1780, 1997. 3, 7, 8

[18] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 5

[19] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 5

[20] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM TOG*, 37:163, 2018. 6

[21] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NeurIPS*, 2014. 5

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint*, arXiv:1312.6114, 2013. 4

[23] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint*, arXiv:1812.08685, 2018. 2, 6

[24] Paweł Korus and Jiwu Huang. Multi-scale analysis strategies in prnu-based tampering localization. *TIFS*, 12:809–824, 2016. 2

[25] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint*, arXiv:1811.00656, 2018. 3

[26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint*, 2019. 1, 2, 3, 6, 7, 8

[27] Kwan-Yee Lin and Guangxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *CVPR*, 2018. 6

[28] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *WACVW*, 2019. 3

[29] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 4

[30] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint*, arXiv:1906.06876, 2019. 3

[31] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. 6

[32] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *WIFS*, 2017. 3

[33] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *arXiv preprint*, arXiv:1901.08971, 2019. 1, 2, 3, 6, 7, 8

[34] Conrad Sanderson. The vidtimit database. Technical report, IDIAP, 2002. 2

[35] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *arXiv preprint*, arXiv:1904.12356, 2019. 2

[36] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 2

[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 3, 7, 8

[38] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 3, 7, 8

[39] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint*, arXiv:1808.06601, 2018. 6

[40] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 4

[41] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*, pages 8261–8265, 2019. 2, 3, 6

[42] Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. Detecting image splicing in the wild (web). In *ICMEW*, 2015. 2

[43] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *CVPRW*, 2017. 3