VIDEOFORENSICSHQ: DETECTING HIGH-QUALITY MANIPULATED FACE VIDEOS

Gereon Fox, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel,
Mohamed Elgharib & Christian Theobalt

Max Planck Institute for Informatics, Saarland Informatics Campus
{gfox,wliu,hyeongwoo.kim,hpseidel,elgharib,theobalt}@mpi-inf.mpg.de

ABSTRACT

There are concerns that new approaches to the synthesis of high quality face videos may be misused to manipulate videos with malicious intent. The research community therefore developed methods for the detection of modified footage and assembled benchmark datasets for this task. In this paper, we examine how the performance of forgery detectors depends on the presence of artefacts that the human eye can see. We introduce a new benchmark dataset for face video forgery detection, of unprecedented quality. It allows us to demonstrate that existing detection techniques have difficulties detecting fakes that reliably fool the human eye. We thus introduce a new family of detectors that examine combinations of spatial and temporal features and outperform existing approaches both in terms of detection accuracy and generalization.

Index Terms— forgery detection, dataset, detectors

1. INTRODUCTION

Model- and learning-based approaches to face video synthesis have reached high levels of visual realism. Some allow facial expressions to be modified or transferred [1–3], while others implement face swapping, i.e. replacing the face interior with a different face identity [4]. Reacting to concerns that these could be misused to modify videos in unethical ways, the research community has developed techniques to detect forgeries, for generic content [5–7] as well as specifically for faces [8–10]. In order to compare the effectiveness of forgery detection methods it is vital to evaluate them on benchmark datasets. As one example, FaceForensics++ [8] contains internet videos modified by several face synthesis techniques [1, 3, 11–13] and demonstrates that an off-the-shelf image classifier, XceptionNet [14], outperforms methods specifically designed for fake detection. However, whenever a forgery detector achieves a high detection accuracy on a dataset, we must wonder: Does this mean that the detector is very good, or does it mean that the fakes in the dataset are just too easy to detect? Based on the observation that the fakes in existing benchmark datasets of forged face videos seem to be easy to spot for the *human* eye (Figure 2), we have formulated the following hypothesis:

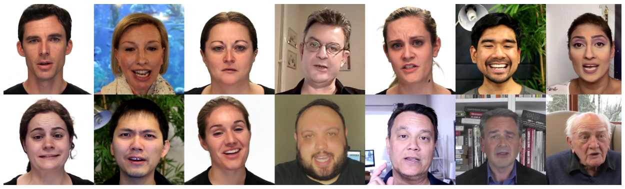


Fig. 1. Forgeries from VideoForensicsHQ.

Hypothesis (H): *The accuracy of existing face video forgery detection methods depends on visual artefacts that humans would be able to spot with the naked eye. As soon as fakes are missing such artefacts, detector performance will drop.*

The artefacts in question include temporal jitter, implausible lighting, unnatural smoothness and blending boundaries, occuring as part of the synthesis process. In the course of investigating **H** we make two main contributions:

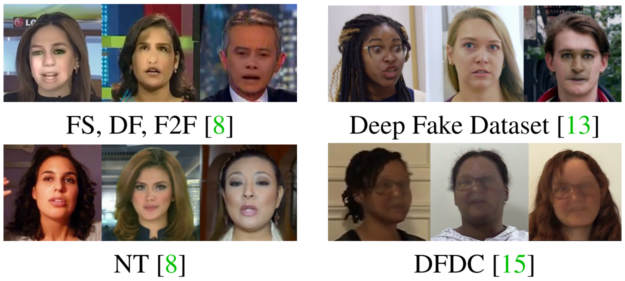
First, we present VideoForensicsHQ, a benchmark dataset of high quality face video manipulations, designed to *not* include said artefacts (Figure 1). Our user study shows (see supplemental) that humans find our fakes considerably harder to detect than in previous datasets. Only VideoForensicsHQ allows us to investigate **H**, by evaluating existing detectors on it, showing that their performance leaves room for improvement. Second, making use of this room, we present a novel family of learning-based detectors that examine combinations of color, low-level noise and temporal correlations. We find these to perform better than previous methods on high-quality fakes and to even generalize to unseen synthesis methods.

2. RELATED WORK

Face Reenactment & Editing: Face reenactment techniques control facial expressions in a video [1, 2, 16, 17]. Many of them extract expressions by fitting a face model [18] and then re-synthesizing the face with parameters copied from a source video [1, 2, 16]. Kim *et al.* [2] for the first time showed space-time coherent realistic global pose and expression editing in videos using a GAN. [16] is a more comprehensive overview.

Face Manipulation Datasets: Several datasets of manipu-

978-1-6654-3864-3/21/$31.00 ©2021 IEEE



FS, DF, F2F [8]

Deep Fake Dataset [13]

NT [8]

DFDC [15]

Fig. 2. Previous face video manipulation datasets contain noticeable artefacts. "Neural Textures" [8] offers the best quality so far, which is why we included it in our user study.

lated images [19, 20] or videos [8, 15, 20] exist. Roessler *et al.*'s FaceForensics++ dataset [8] contains 1,000 videos, each manipulated by 4 different techniques [1, 3, 11, 12]. Their results show that many manipulation approaches produce very noticeable artefacts. Google released the Deep Fake Detection Dataset [13]. Many sequences exhibit visual artefacts. Facebook released a dataset [21] of manipulated face videos of varying quality, with face resolution often much less than 299^2. Jiang *et al.* presented DeeperForensics 1.0 [22], which, augmentations aside, provides 1000 forgeries derived from [8]. As can be seen in Figure 2, our user study, and our evaluation (Section 5), all these datasets are made up of fakes that are easily detected by humans and machines.

Detection of Manipulated Visual Content: Generic detection techniques [5, 7, 23–27] often examine low-level features such as high-frequency components and noise. Fridrich *et al.* [23] introduced convolutional kernels designed for steganalysis, that were later formulated as a constrained CNN [24]. Bayar *et al.* [5] suppress image content to focus on low-level patterns. The work of [27] localizes edited regions of an image by examining so-called "noiseprints". Face-specific techniques can be classified into single-image-based [6–9, 19, 28–30] and multi-image-based [10, 31] approaches. Zhou *et al.* [19] detect face swaps using a two stream network (visual artefacts + steganalysis features). *MesoInc-4* [6] is a CNN that learns at which level of granularity to investigate the input. Rössler *et al.* [8] examined a variety of manipulation detection techniques [6, 23, 24] on their FaceForensics++ dataset, with XceptionNet [14] emerging as the most robust detector. Other works have studied temporal correlations, based on "action-units" [9] or RNN's [10].

3. THE VideoForensicsHQ DATASET

To investigate **H**, we need a benchmark dataset that contains *many* fakes of high quality: In order for humans to be unable to spot fakes, we must avoid artefacts such as temporal jitter, unnatural movement, implausible lighting, unusual smoothness, or strong blending boundaries. While there definitely are state-of-the-art synthesis techniques that achieve such quality under ideal conditions, we are not aware of a *large-scale* benchmark dataset that aggregates *many* such high quality results. VideoForensicsHQ is the first such dataset, as confirmed by our user study (supplemental material). The challenge does *not* lie in finding a novel synthesis method and it is by no means our goal to present one. Instead we adapt Deep Video Portraits (DVP) [2] for *large-scale* fake creation.

While DVP can transfer performances from a source person to a *different* target person, this mode can lead to artefacts if the distribution of facial expressions differs a lot between source and target. Not even the "style-preserving" variant [16] avoids glitches as reliably as necessary. We thus produce "intra-person" transfers (i.e. source and target are the same person). Recent works [32, 33] show this to be a very relevant threat-scenario. Since DVP is trained on a set of frames that is disjoint from the source/target sequence, it has not seen the expressions to synthesize in advance and must still generate the typical GAN artefacts that are common with synthesis methods, but typically go unnoticed by humans.

VideoForensicsHQ At-A-Glance: VideoForensicsHQ contains 1737 videos of talking faces (43% male, 57% female), with 8 different emotions. Most videos have resolution 1280×720. They amount to 1,666,816 frames with average resolution 968^2 and the average face covering 487^2 pixels. There are three different subsets: Group#1 was mined from [16], Group#2 from RAVDESS [34], and Group#3 from YouTube. In total, our dataset contains 326,973 fake frames, comparable to the "Neural Textures" [3] part of FaceForesnics++. While their fakes are the ones that come closest to our dataset in terms of visual quality (see Figure 2), our user study (supplemental) shows that our fakes are much harder to detect for humans: 65.8% of our fakes are mistaken as reals, while only 14.3% of the "Neural Textures" fakes pass this test. For more details see our supplemental document.

Production process: Mining real videos as the basis for our fakes is challenging because jump-cuts, animations and unusual face poses need to be circumvented automatically, especially for YouTube. To synthesize video of an identity, DVP requires about 5 - 10 minutes of training material, with all frames showing the same face at roughly the same distance, in a near-frontal pose. To find such material, we run a facial landmark tracker [35] on all frames of all source videos, obtaining 66 landmark positions for every frame f_i, and one confidence value in the range $[0, 1]$ for every landmark position. We compute three metrics:

1. c_i: avg. landmark confidence for frame f_i
2. d_i: avg. offset between landmark positions in f_i and f_{i-1}, divided by face size
3. mean and standard deviation of the c_i's and d_i's

A frame is regarded unsuitable in any of the following cases: a) $c_i < 0.2$, b) $d_i > 0.1$, c) $c_i < 0.6$ deviates from the confi-

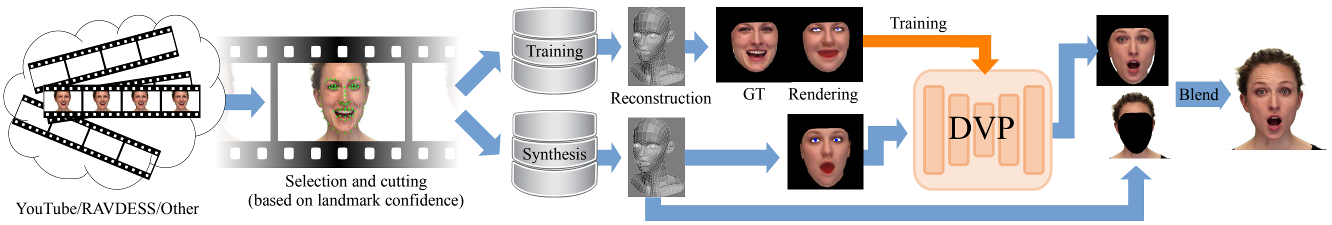


Fig. 3. We apply monocular reconstruction to videos from the "Synthesis" set, to obtain facial parameters that are close to the training distribution. DVP [2] turns these into photorealistic videos.

dence mean by more than 110% of the std.dev. (in neg. direction), d) $d_i > 0.025$ deviates from the displacement mean by more than 110% of the std.dev. (in pos. direction). If none of these apply, the frame is added to the current segment of suitable frames. A *segment* here is a contiguous set of frames, with no scene cuts. We add the longest good segments to the *training* set for the respective identity until 5000 to 6000 frames are reached. All good segments beyond that make up the disjoint *synthesis* set for this identity.

The *training* set is processed with a monocular face reconstruction approach [18], which encodes the facial performance as a sequence of parameter vectors. The vectors are then rendered to obtain the conditioning input that DVP learns to turn into RGB output again (Figure 3). Thus for each identity, we obtain one DVP model that can render facial performances at photorealistic quality. The input to such a model can be any arbitrary facial performance, also given as a sequence of parameter vectors. But to reliably avoid strong artefacts, we need to give facial performances as input that are close to the distribution that DVP saw during training (without, of course, using any of the training data!). We simulate a faker that is able to synthesize such parameter sequences, by applying monocular reconstruction to the *synthesis* set as well, thereby obtaining parameters that have the necessary properties. This is why our fakes mostly avoid clearly visible glitches, but still preserve the less noticeable artefacts that every GAN-based synthesis method inevitably exhibits.

For further details, including our modifications to DVP, we refer to our supplemental document. More synthesis results are shown in our submission video.

4. DETECTING HIGH-QUALITY FACE MANIPULATIONS

We consider XceptionNet [14] a representative of existing face video forgery detection methods, because it ranked highest in FaceForensics++ [8]. If **H** is true, XceptionNet should perform worse on VideoForensicsHQ than it does on FaceForensics++. This expectation is justified because XceptionNet is a generic image classifier that has not been designed for fake detection and thus should look for clearly visible artefacts in the image space. Since it nevertheless outperformed all other, detection-specific methods in [8], we want to enhance its ability to detect seemingly flawless fakes, without compromising its ability to detect strong artefacts. We thus present a novel family of detectors that examine *combinations* of multiple cues (Figure 4): the original RGB values, low-level spatial noise, and temporal correlations.

XceptionNet consists of an entry flow $\mathrm{In}_{d\alpha\beta\gamma\delta\epsilon}$, a middle flow M, and an exit flow Out. Parameters α, β, γ, δ, and ϵ specify the number of features per convolutional layer (see supplemental), while d specifies the number of input channels. One can denote XceptionNet as the function $C := \mathrm{In}_{3,32,64,128,256,728} \circ \mathrm{M}^8 \circ \mathrm{Out}$, applied to color images, yielding a score for class "real" and one for "fake". In the following, leading or trailing zeros in the indices of $\mathrm{In}_{d,\alpha,\beta,\gamma,\delta,\epsilon}$ disable the respective layers.

Repetitions of M drive up memory consumption and training overhead. To test whether 8 repetitions are actually necessary for forgery detection, we remove M entirely:

$$\mathrm{B} := \mathrm{In}_{3,32,64,128,256,728} \circ \mathrm{M}^0 \circ \mathrm{Out}$$

Since VideoForensicsHQ contains very few strong visual artefacts that C or B could easily pick up, we define S to not classify frames $F \in \mathbb{R}^{299\times299\times3}$ themselves, but their spatially high-pass-filtered versions $\frac{1}{2}\cdot(F - g * F) + \frac{1}{2}$, where g is a Gaussian kernel of size 5, with standard deviation $\sigma = 1.1$. The architecture of S is that of B.

Our combination of C and S,

$$\mathrm{CS} := (\mathrm{In}_{3,32,64,128,256,364}, \mathrm{In}_{3,32,64,128,256,364}) \circ \mathrm{M}^2 \circ \mathrm{Out}$$

receives the same inputs as C and S and fuses the color and noise features just before entering M^2, where the combined receptive field of the convolutional kernels has size 17×17 (see Figure 4). We extend CS to

$$\begin{aligned}
\mathrm{PP}_{\mathrm{ct}} &:= (\mathrm{In}_{3,32,64,0,0,0}, \mathrm{In}_{3,8,8,0,0,0}) \circ \mathrm{In}_{0,0,72,128,256,512} \\
\mathrm{PP}_{\mathrm{cts}} &:= (\mathrm{PP}_{\mathrm{ct}}, \mathrm{In}_{3,16,32,64,128,256}) \\
\mathrm{CST} &:= \mathrm{PP}_{\mathrm{cts}} \circ \mathrm{M}^1 \circ \mathrm{Out}
\end{aligned}$$

which receives temporal features as a third input (Figure 4). They are extracted as follows:

1. Spatial Gaussian kernel (size 49, $\sigma = 7.7$), suppressing high spatial frequencies that motion would turn into temporal ones (e.g. an edge sweeping over a pixel).

2. Pixel-wise temporal high-pass filtering of the form $A_i := -\frac{1}{4}F_{i-1} + \frac{1}{2}F_i + -\frac{1}{4}F_{i+1}$.
3. Batch normalization.
4. Amplitudes smaller than t are dampened by computing $A'_i := \mathrm{thr}_t(A_i) - \mathrm{thr}_t(-A_i)$, where the function $\mathrm{thr}_t(A) := \frac{t}{10} \cdot (\ln(1 + \exp(x)) + 10 \cdot \mathrm{sigmoid}(x))$ for $x := \frac{10}{t} \cdot (A - t)$ is smooth and differentiable in t.
5. Computation of temporal gradients: $G_i := |A_i - A_{i-1}|$.
6. Temporal lowpass filtering (kernel $(\frac{1}{32}, \frac{1}{8}, \frac{3}{16}, \frac{1}{8}, \frac{1}{32})$).

This process emphasizes unnaturally fast motions (i.e. quick changes from one frame to the next), often observed in forgeries. With the exception of step 1, all operations are pixel-wise. Step 2 suppresses low temporal frequencies, which are likely of natural origin. Steps 3 and 4 ensure that among the high frequency spikes we focus on those of a certain minimum amplitude, which are most likely artificial. Step 5 turns oscillations between + and - into large positive values. Step 6 stabilizes the resulting signals, such that more output frames exhibit bright regions that the classifier can detect. For more information on thr_t see our supplemental material.

We are not aware of any existing face video forgery detectors that use such an approach. We deliberately resist any reflex to make "everything trainable" in our feature extraction, to prevent it from overfitting to the training data. Only threshold t is trainable, and our evaluation (Section 5.2) shows that our detectors generalize better than completely trainable ones.

The number of repetitions of M and the points at which we fuse streams were empirically chosen to maximize detection accuracy while not exceeding the 11GB of GPU memory in an NVIDIA 1080Ti. The resulting tradeoffs can lead to S performing slightly better than CS and CST on videos in which color and temporal information do not give a benefit over spatial noise, because the latter two cannot dedicate as much memory to spatial noise as S (Section 5). On the other hand, spatial noise cues alone do not generalize as well as combinations with other types of information (Section 5.2).

5. RESULTS

Based on our new dataset VideoForensicsHQ and our novel family of detectors we can now investigate **H**:

State of the Art Techniques We compare our detectors to a number of related techniques: Detector C, published as "XceptionNet" [14], performs best in the FaceForensics++ benchmark [8]. We evaluate *MesoInc-4* [6], *Bayar et al.* [5] and *Durall et al.* [30] as they show good results in analysing low-level features. We also compare to *Wang et al.* [7], a recent classifier that generalizes to unseen rendering methods.

Preprocessings and training Videos were preprocessed with one common pipeline before being fed into all detectors (temporally smooth face crops, rescaling to appropriate resolutions). Imbalances in the datasets were accounted for by weighted sampling. More details in supplementary material.

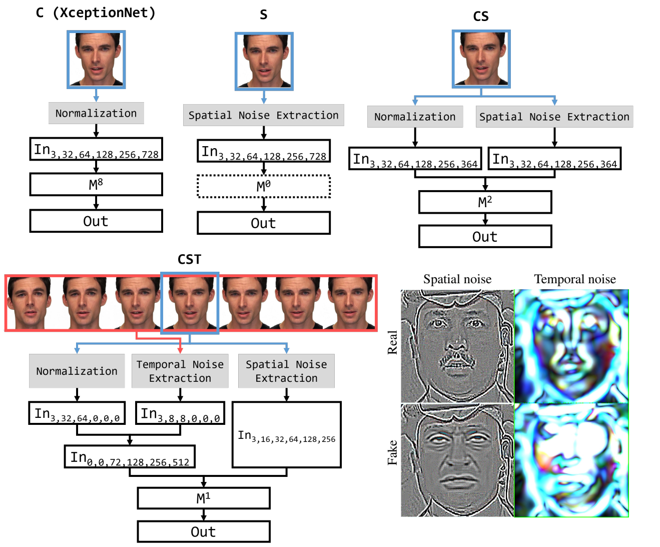


Fig. 4. Our detectors extend XceptionNet [14] to a multi-stream classifier for combinations of color, spatial noise and even temporal features.

5.1. Detecting highly photorealistic manipulations

To test **H**, we trained all detectors on FaceForensics++ [8] and DeeperForensics 1.0 [22], which both contain strong visual artefacts (Figure 2), as well as on VideoForensicsHQ, which does not (Figure 1). Table 1 confirms **H**: With the exception of our novel detectors all accuracies are considerably lower for VideoForensicsHQ, than for previous datasets. In fact, XceptionNet (C), the best detector in [8] drops by more than 10% and is even outperformed by B, which is a reduced version of C. Our detectors S, CS and CST on the other hand perform well on all three datasets. We observe that CS and CST perform not quite as well as S. This is because they had to sacrifice some of the GPU memory that S can dedicate to spatial noise, in order to handle color and temporal features (Section 4). Since VideoForensicsHQ does not contain strong visual or temporal artefacts, this sacrifice does not pay off.

We remark that only our dataset is able to differentiate the best detectors from one another, while on previous datasets many detectors achieve close to 100% accuracy.

5.2. Generalization across manipulation techniques

Since the synthesis method for a forgery is often unknown, detectors should generalize to unseen methods.

To evaluate this ability, we train detectors on the Face-Forensics++ subsets $FS \cup NT$ (FaceSwap + Neural Textures [3]) and $F2F \cup DF$ (Face2Face [1] + Deep Fakes) and then test them on the subset they were *not* trained on. We

Table 1. Test accuracies and validation accuracy maxima of our detectors (B, S, CS, CST) and previous approaches.

Arch.	FF++	Deeper Forensics	VFHQ
C	99.23%	**98.64%**	88.59%
B	99.34%	98.09%	91.95%
S	**99.38%**	98.21%	**99.45%**
CS	99.25%	98.32%	97.12%
CST	99.35%	98.34%	97.78%
MesoInc-4	92.44%	97.25%	76.73%
Wang et al.	75.55%	61.05%	56.44%
Durall et al.	56.69%	62.88%	61.98%
Bayar et al.	95.82%	97.89%	74.65%

Note: Numbers averaged over 2 training runs.

Table 2. Training on FaceForensics++ subsets $FS \cup NT$ and $F2F \cup DF$, with test accuracies for the opposite subsets.

Arch.	Train: $FS \cup NT$		Train: $F2F \cup DF$	
	Acc.F2F	Acc.DF	Acc.NT	Acc.FS
C	82.25%	92.91%	58.40%	50.09%
B	90.40%	95.15%	66.03%	50.27%
S	98.46%	94.93%	85.21%	55.19%
CS	**99.46%**	98.74%	86.74%	51.78%
CST	98.91%	**99.03%**	**90.65%**	56.77%
MesoInc-4	89.65%	71.94%	73.10%	49.76%
Wang et al.	77.91%	80.03%	84.40%	**58.89%**
Durall et al.	55.87%	55.47%	53.68%	54.19%
Bayar et al.	64.22%	94.53%	64.04%	50.02%

Note: Numbers averaged over 3 training runs.

do not use VideoForensicsHQ for this experiment, because it contains only one manipulation technique and differs from other datasets in more respects than just the synthesis method (much higher resolution faces, no visible cues, etc). FaceForensics++ is better suited here because all other factors can be kept constant when switching between synthesis methods.

Table 2 shows that training our detectors on $FS \cup NT$ makes them generalize well to $F2F \cup DF$, where they outperform existing methods. CS ranking higher than S and C suggests that *combining* color and spatial noise can help generalization. The inverse, (train on $F2F \cup DF$, test on FS), gives low accuracies for all detectors, suggesting that FS contains artefacts not seen in $F2F \cup DF$. The subset NT seems to be easier to generalize to, with our detectors tending to outperform existing techniques and CST ranking highest.

Although detectors like C or *Bayar et al.* could theoretically learn the spatial filtering we hardcoded for S, we see them perform considerably worse than S in Tables 1 and 2.

5.3. Further results

Since no single dataset can cover all variations of forged video content (synthesis methods, image qualities, lighting conditions, camera angles, etc.), a robust detector should support training on a *union* of datasets. We evaluated detectors on such a union and found ours to outperform previous ones, with CST ranking highest. We also investigated the impact of the number of training identities and found our detectors to require fewer identities than previous methods. Detailed results can be found in our supplemental document.

6. CONCLUSION

In this paper, we have introduced VideoForensicsHQ, the first benchmark set for face video detection that provides a large number of manipulations a human would not be able to spot. Only with this dataset were we able to investigate whether current approaches to face video forgery detection are ready for the advent of synthesis methods that produce seemingly "perfect" results, confirming hypothesis **H**. To compensate for the shortcomings of existing detection approaches in this scenario, we have introduced a novel family of detectors that combine spatial and temporal information in a way that has not been used in the area of face video forgery detection before. We have shown our detectors to outperform related methods both on previous datasets and on VideoForensicsHQ.

While at first sight one might mistake the "intra-person" expression transfers in our dataset as harmless, recent works [32, 33] demonstrate that even slight manipulations of this kind can have dramatic consequences. VideoForensicsHQ is the first and so far only benchmark dataset that allows their study. The absence of human-detectable artefacts in VideoForensicsHQ has the advantage of preventing detectors from learning to rely on their presence. This suggests that VideoForensicsHQ should be included in any "serious" detector training set and allows the detection community to prepare for future advances in forgery approaches already today.

Acknowledgement: This work was supported by the ERC Consolidator Grant 4DReply (770784).

7. REFERENCES

[1] Justus Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time Face Capture and Reenactment of RGB Videos," in *CVPR*, 2016.

[2] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt, "Deep Video Portraits," *ACM TOG*, 2018.

[3] Justus Thies, Michael Zollhöfer, and Matthias Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM TOG*, 2019.

[4] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormaehlen, Patrick Perez, and Christian Theobalt, “Automatic face reenactment,” in *CVPR*, 2014.

[5] B. Bayar and M. C. Stamm, “Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection,” *TIFS*, 2018.

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

[7] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros, “CNN-generated images are surprisingly easy to spot...for now,” in *CVPR*, 2020.

[8] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *ICCV*, 2019.

[9] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li, “Protecting world leaders against deep fakes,” in *CVPRW*, 2019.

[10] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan, “Recurrent convolutional strategies for face manipulation detection in videos,” in *CVPRW*, 2019.

[11] “Deepfakes. GitHub.,” https://github.com/deepfakes/faceswap.

[12] “Faceswap. GitHub.,” https://github.com/MarekKowalski/FaceSwap/.

[13] Nick Dufour and Andrew Gully, “Contributing data to deepfake detection research. Google Blog.,” https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

[14] François Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *CVPR*, 2017.

[15] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer, “The deepfake detection challenge (DFDC) preview dataset,” 2019.

[16] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt, “Neural style-preserving visual dubbing,” *ACM TOG*, 2019.

[17] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky, “Fast bi-layer neural synthesis of one-shot realistic head avatars,” in *ECCV*, 2020.

[18] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt, “Reconstruction of personalized 3d face rigs from monocular video,” *ACM TOG*, 2016.

[19] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis, “Two-stream neural networks for tampered face detection,” in *CVPRW*, 2017.

[20] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus, “Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation,” in *WACVW*, 2019.

[21] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer, “The deepfake detection challenge dataset,” 2020.

[22] 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,” 2020.

[23] Jessica Fridrich and Jan Kodovsky, “Rich models for steganalysis of digital images,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, 2012.

[24] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva, “Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection,” in *ACM IH&MMSec*, 2017.

[25] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, “Learning rich features for image manipulation detection,” in *CVPR*, 2018.

[26] Y. Wu, W. AbdAlmageed, and P. Natarajan, “Mantranet: Manipulation tracing network for detection and localization of image forgeries with anomalous features,” in *CVPR*, 2019.

[27] D. Cozzolino and L. Verdoliva, “Noiseprint: A cnn-based camera model fingerprint,” *TIFS*, 2020.

[28] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo, “Face x-ray for more general face forgery detection.,” *CoRR*, 2019.

[29] Yuezun Li and Siwei Lyu, “Exposing deepfake videos by detecting face warping artifacts,” in *CVPRW*, 2019.

[30] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper, “Unmasking deepfakes with simple features,” 2019.

[31] Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu, “Towards generalizable forgery detection with locality-aware autoencoder,” 2019.

[32] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman, “Synthesizing Obama: learning lip sync from audio.,” *ACM TOG*, 2017.

[33] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala, “Text-based editing of talking-head video,” *ACM TOG*, 2019.

[34] Steven R Livingstone and Frank A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english,” in *PloS one*, 2015.

[35] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn, “Deformable model fitting by regularized landmark mean-shift,” *IJCV*, vol. 91, no. 2, 2011.

VIDEOFORENSICSHQ: DETECTING HIGH-QUALITY MANIPULATED FACE VIDEOS – SUPPLEMENTARY MATERIAL –

Gereon Fox, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel,
Mohamed Elgharib & Christian Theobalt

Max Planck Institute for Informatics, Saarland Informatics Campus
{gfox,wliu,hyeongwoo.kim,hpseidel,elgharib,theobalt}@mpi-inf.mpg.de

Table 1. VideoForensicsHQ subsets

Subset	Fake frames	Real frames
group#1	60,058	119,992
group#2	74,765	190,259
group#3	192,150	1,029,592
total	32,6973	1,339,843

Table 2. User study results

Source	Rated “fake”	Rated “real”
Real videos	15.0%	85.0%
“Neural Textures” fakes [4]	85.7%	14.3%
VideoForensicsHQ fakes	34.2%	65.8%

1. SYNTHESIS DETAILS

We mined the authentic raw footage for our dataset from [1] (Group#1), RAVDESS [2] (Group#2) and YouTube (Group#3). Table 1 lists the sizes of these subsets.

The original version of DVP [3] cannot handle dynamic backgrounds and works at a fixed resolution of 256^2. We thus prepare training frames by cropping them around the face and masking out the background (see pipeline figure in the main paper), in order to make the network focus all its capacity on the face region. The resulting square images are scaled to resolution 256^2. Instead of a separate conditioning image for the eye-gaze (like in [3]) we overlay the eye gaze rendering on the face rendering. We use temporal supervision for DVP, by means of “temporal windows”, as proposed in [3]: The discriminator sees temporal volumes of 5 frames for most of Group#2. We found the temporal window to not improve quality considerably, which is why Group#3 and Group#1 were synthesized with window size 1. We train our DVP models for up to 200 epochs, estimating the mean squared photometric error against ground truth on the validation set. The model with the smallest error is used for synthesis. Since the facial performances we are rendering at synthesis time have been reconstructed from real footage, we know the coordinates of the face region in that footage. This allows us to alpha-blend the DVP output into those original frames.

Figure 1 shows more forgery examples from our dataset.

2. USER STUDY

We conducted a user study to asses the quality of our fakes compared to FaceForensics++ [4]. We randomly selected 13 manipulated videos from VideoForensicsHQ and 13 manipulated videos from the “Neural Textures” subset of FaceForensics++ [4], created with the reenactment technique by Thies *et al.* [5]. Other approaches in FaceForensics++ produce fakes with much more visible artefacts (see Figure 2 in the main manuscript). In addition, we randomly selected 6 unmodified videos from VideoForensicsHQ and 7 from FaceForensics++.

In total our study contains 39 videos, randomly shuffled for each participant. For each video, we recorded the answer to the question “Does the video look real or fake?”. Most participants were computer scientists, with little-to-no knowledge of face manipulation techniques. 61 subjects participated in the study. On average, fakes from VideoForensicsHQ were rated real 65.8% of the time, and fakes from FaceForensics++ were rated real only 14.3% of the time. Table 2 lists the full results. We note that unmodified videos were also rated as manipulated 15% of the time, which reflects a baseline error level in human detection performance. We also asked participants what made them flag a video as modified. Some of the most common responses were:

1. Various visual artefacts, especially in mouth interior
2. Non-natural eye movement
3. Body movements or hand gestures not matching speech
4. Non-natural mouth-related movements e.g. lips being tight when they should not be, deforming/dislodging jaw, etc.
5. Incorrect audio-lip synchronization
6. A single glitch occurring over 2-3 seconds
7. Spoken language not matching language of written text

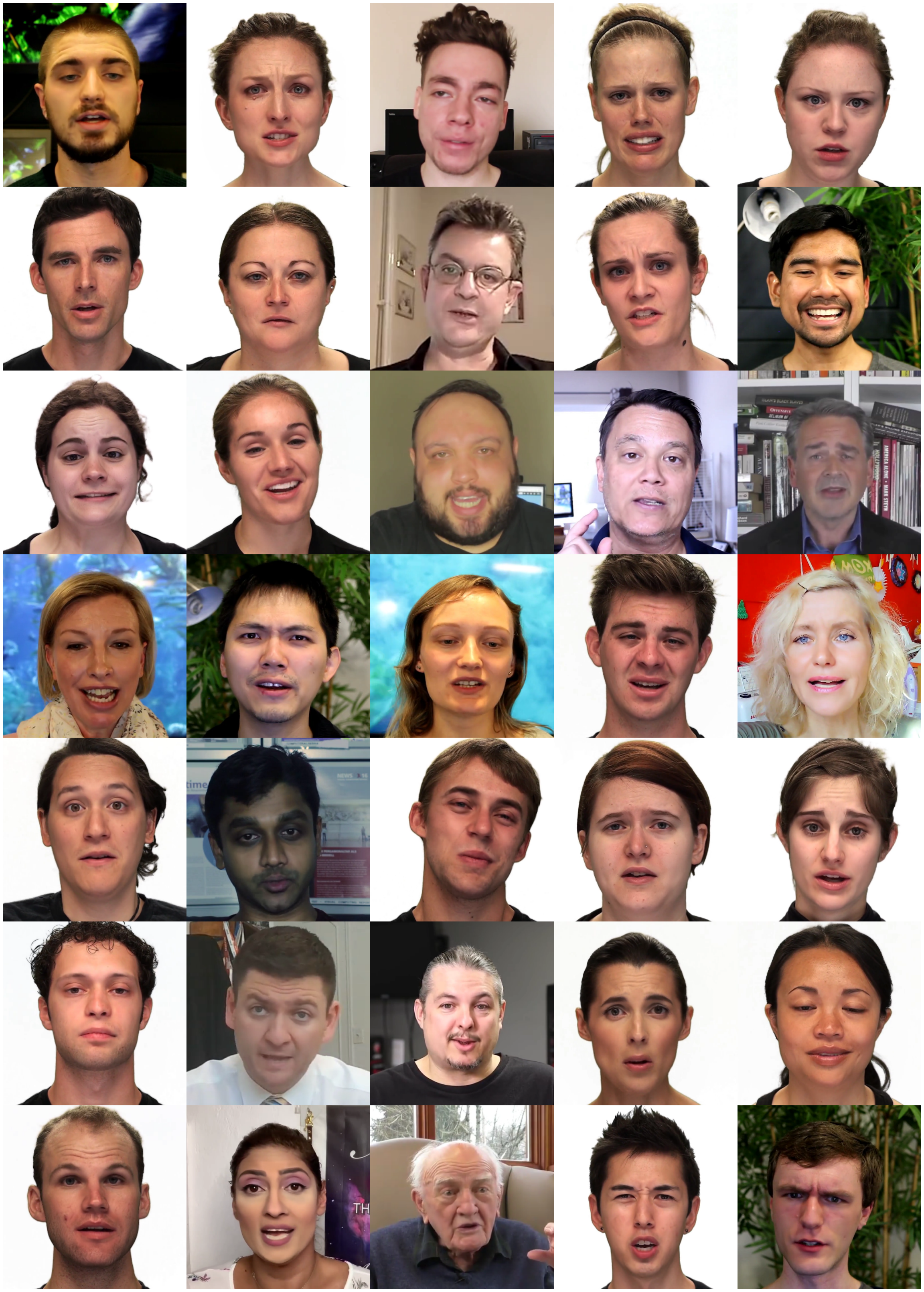


Fig. 1. Forgeries from VideoForensicsHQ.

3. DETECTION DETAILS

3.1. Parametrization of XceptionNet

In order to build our detectors, we have generalized XceptionNet [6], by parametrizing various dimensions of its architecture (see main paper). Figure 2 clarifies the meaning of each parameter we introduced.

3.2. Temporal filtering

The function thr_t is supposed to dampen low amplitudes of a signal. The parameter t is a threshold specifying which amplitudes are to be dampened. Figure 3 plots thr_t for different values of t. The function is differentiable in t, such that the training process can automatically determine a good position and shape for the "cliff".

3.3. Detection preprocessing

All training and test data for all detectors is preprocessed by one common pipeline: Face bounding boxes are computed using *dlib* [7], with temporal smoothing of their coordinates. We extract constant-size square bounding boxes, scaled to resolution 299^2. We resample videos at 25fps. Frames for which no face bounding box can be found are omitted. For *MesoInc-4* we scale the resulting frames to 256^2, whereas for *Durall et al.* we compute a 209-dimensional feature vector as specified in [8].

3.4. Detection training

The Xception-based detectors C, B, S, CS and CST are all trained with stochastic gradient descent (momentum 0.9, weight decay 10^{-5}), multiplying the initial learning rate of 0.03 with factor $0.97^{0.1}$ per epoch. For CST we initialize t with $\frac{1}{40}$. Previous detectors are trained as specified in their publications, except for *Wang et al.* and *Durall et al.*: *Wang et al.* is claimed to generalize well to unseen rendering methods. We thus use its pretrained weights and merely optimize a threshold on its singular output value, based on the ROC curve over the samples that were seen within one epoch of training. We perform this optimization for 5 epochs and average the 5 resulting thresholds. For every training batch of *Durall et al.* [8] we optimize a new SVM model on the Fourier features. At validation and test time we average the predictions of all SVM models obtained in this way.

All detectors are trained with batch size 24, except for *MesoInc-4* (512) , *Durall et al.* (512) and *Bayar et al.* (256). Except for *Wang et al.*, all methods are trained with a hard limit of 100 epochs. We stop training earlier if 5 epochs with a validation accuracy of more than 99% have been seen (not necessarily consecutively). The model with maximal validation accuracy is used at test time.

To account for imbalances in the datasets, we randomly sample 10% of the training frames and 20% of the validation frames in every epoch. Sampling here means to first uniformly select a class ("real"/"fake"), then a subset (which is relevant for VideoForensicsHQ because it consists of three different groups), then a subject and then one of the sequences for this subject. Frames are sampled uniformly from sequences. Since *Durall et al.* is not designed for the amounts of data resulting from the aforementioned sampling rates we lower them to 0.5% training and 1% validation samples for this method.

This training procedure is the reason why in Table 1 of the main paper, the accuracies we report for *MesoInc-4* and *Bayar et al.* on FaceForensics++ are slightly lower than the ones reported in [4]. However, the detectors based on XceptionNet [6] (C, B, S, CS, CST), do not seem to be impacted by this, which we interpret as a strength of XceptionNet-based architectures.

At test time, we evaluate *all* frames of the test set in which a face could be found, but weigh per-frame predictions by the probability of a frame being sampled according to above sampling process.

4. FURTHER RESULTS

4.1. Training on a union of datasets

Since no single dataset can cover all variations of forged video content (synthesis methods, image qualities, lighting conditions, camera angles, etc.), a robust detector should support training on a *union* of datasets. To evaluate how well detectors handle this setup we have trained them on the union of FaceForensics++, VideoForensicsHQ and DFDC (preview) [9] and tested them on FaceForensics++ and VideoForensicsHQ (Table 3[1]). Compared to training on only one single dataset (see main paper), the task is now hard enough to also differentiate *our* detectors from one another: B again performs better than C. S and CS are on par. CST can now demonstrate the benefit of temporal information, ranking highest on both test sets.

4.2. Impact of training corpus size

VideoForensicsHQ contains only 45 identities, while FaceForensics++ contains 1000 identities. This raises the question of how many identities are necessary to train a detector.

We have thus randomly sampled small training sets from VideoForensicsHQ, with different numbers of identities. Detectors were trained on these subsets and then tested on random test sets of 15 identities each (disjoint from the training sets). Figure 4 shows the average accuracies after training on

[1] [9] contains very challenging perspectives and lighting conditions, as well as fast motion, making our preprocessing struggle to the point that it becomes the limiting factor of accuracy: All detectors, ours and previous, achieve about 80% test accuracy.

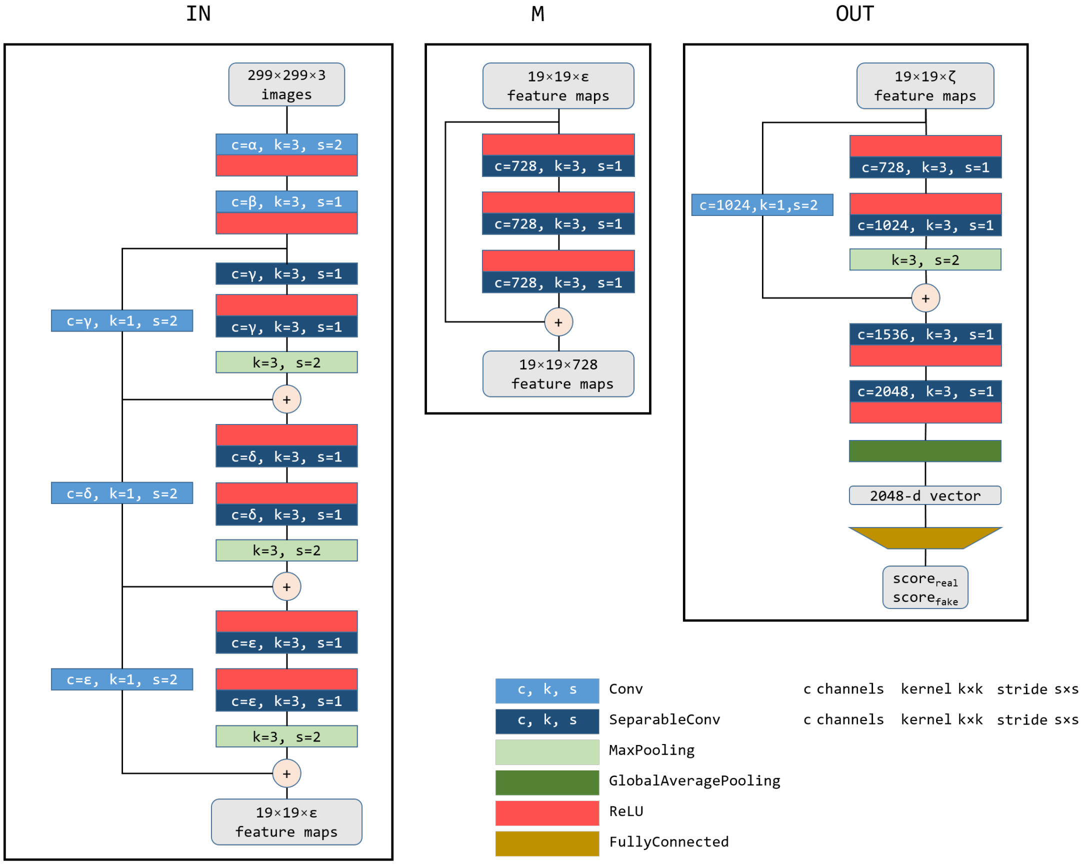


Fig. 2. The building blocks of XceptionNet [6] are the basis of our multi-stream detectors (see Figure 4 of the main manuscript). In order to trade memory capacity between multiple streams and fuse them at the right locations, we changed the numbers of features in each layer of each instance of such a building block. ϵ in M and ζ in Out are determined by the number of output feature in the preceding block.

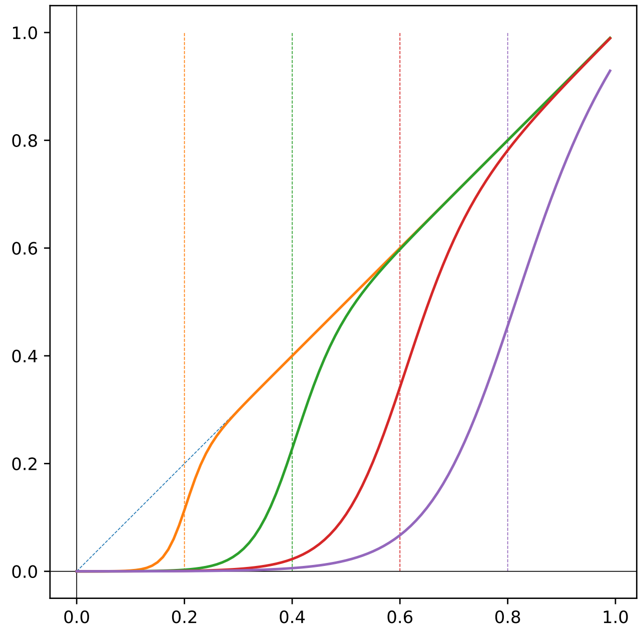


Fig. 3. Graphs of thr_t, for $t \in \{0.2, 0.4, 0.6, 0.8\}$.

different numbers of identities. For each number of identities we have sampled 3 to 5 different training sets.

We observe that the best detectors achieve close to 100% test accuracy already for training corpora of only 25 identities (training + validation), which is much fewer than the total number of identities in VideoForensicsHQ, providing evidence that our dataset is sufficient to generalize to unseen identities, and that our detectors do *not* overfit to the training identities.

5. REFERENCES

[1] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt, "Neural style-preserving visual dubbing," *ACM TOG*, 2019.

[2] Steven R Livingstone and Frank A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," in *PloS one*, 2015.

[3] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt, "Deep Video Portraits," *ACM TOG*, 2018.

[4] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV*, 2019.

[5] Justus Thies, Michael Zollhöfer, and Matthias Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM TOG*, 2019.

Table 3. Test accuracies after training on the union of FaceForensics++, VideoForensicsHQ and DFDC (preview) [9].

Arch.	Test accuracy	
	FaceForensics++	VideoForensicsHQ
C	95.90%	78.99%
B	96.16%	80.02%
S	97.69%	88.94%
CS	98.01%	87.91%
CST	**98.67%**	**90.63%**
MesoInc-4	74.34%	76.65%
Wang et al.	75.88%	56.75%
Durall et al.	54.20%	55.51%
Bayar et al.	90.02%	78.85%

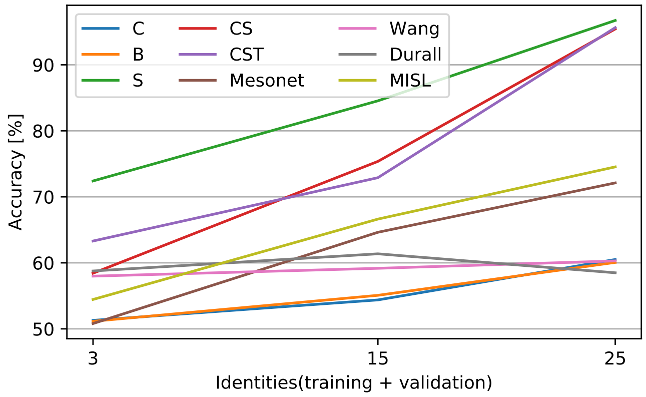


Fig. 4. Average test accuracies achieved by models that were trained on VideoForensicsHQ subsets containing different numbers of identities, see Section 4.2.

[6] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *CVPR*, 2017.

[7] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, 2009.

[8] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper, "Unmasking deepfakes with simple features," 2019.

[9] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019.