

PROTECTING PRESIDENT ZELENSKYY AGAINST DEEP FAKES

A PREPRINT

Matyáš Boháček

Gymnasium of Johannes Kepler
Parléřova 2/118, 169 00 Praha 6, Czech Republic
matyas.bohacek@matsworld.io

Hany Farid

Department of Electrical Engineering and Computer Sciences
School of Information
University of California, Berkeley
hfarid@berkeley.edu

ABSTRACT

The 2022 Russian invasion of Ukraine is being fought on two fronts: a brutal ground war and a duplicitous disinformation campaign designed to conceal and justify Russia’s actions. This campaign includes at least one example of a deep-fake video purportedly showing Ukrainian President Zelenskyy admitting defeat and surrendering. In anticipation of future attacks of this form, we describe a facial and gestural behavioral model that captures distinctive characteristics of Zelenskyy’s speaking style. Trained on over eight hours of authentic video from four different settings, we show that this behavioral model can distinguish Zelenskyy from deep-fake imitators. This model can play an important role – particularly during the fog of war – in distinguishing the real from the fake.

Keywords Deep fakes · Disinformation · Digital Forensics · Facial Mannerisms · Gestural Mannerisms

1 Introduction

In the early days of the Russian invasion of Ukraine, President Zelenskyy warned the world that Russia’s digital disinformation machinery would create a deep fake of him admitting defeat and surrendering. A few weeks later in mid-March of 2022, a deep fake of Zelenskyy appeared with just this message [Allyn, 2022]. This video, Figure 1, was quickly debunked thanks to the rather crude audio and video and to Zelenskyy’s pre-bunking. This type of deep fake, however, is likely just the beginning of a new digital front that we might expect in this and future conflicts.

A recent set of perceptual studies [Groh et al., 2022] examined the ability of untrained observers to distinguish between real and deep-fake videos. In one condition, participants viewed a single video and categorized it as real or fake. Participants correctly identified 66% of the deep-fake videos, as compared to chance performance of 50% (pooled responses from all participants – so-called crowd wisdom – yields an improved accuracy of 80%). In a second condition,



Figure 1: One video frame of the real (left) and deep-fake version (right) of Ukrainian President Zelenskyy.



Figure 2: Four representative examples of President Zelenskyy in different contexts: (a) public address; (b) press briefing; (c) bunker; and (d) armchair.

participants were shown the prediction by the top-performing DFDC computational model [Ferrer, 2020] and given the opportunity to update their response. In this collaborative condition, individual participant accuracy improved to 73%.

While we may have the ability to perceptually detect some deep-fake videos, our ability is not terribly reliable and this task will become increasingly more difficult as deep fakes continue to improve in quality and sophistication. We must, therefore, turn to computational methods to assist in the task of distinguishing the real from the fake.

The computational detection of deep-fake videos can be partitioned into three basic categories: (1) learning-based, in which features that distinguish real from fake content are explicitly learned by any of a range of different machine-learning techniques [Zhou et al., 2017, Afchar et al., 2018, Li et al., 2019]; (2) artifact-based, in which a range of low-level (pixel based) to high-level (semantic based) features are explicitly designed to distinguish between real and fake content [Li et al., 2018, Agarwal and Farid, 2021, Agarwal et al., 2020a]; and (3) identity-based, in which biometric-style features are used to identify if the person depicted in a video is who it purports to be [Agarwal et al., 2019, 2020b, 2021, Cozzolino et al., 2021].

The advantage of learning-based approaches is they are able to learn detailed and subtle video-synthesis artifacts. The disadvantage is these techniques often struggle to generalize to new content not explicitly part of the training data set, and can be vulnerable to adversarial attacks [Carlini and Farid, 2020], and simple laundering attacks where the synthesized media is trans-coded or resized [Barni et al., 2018]. In the 2019-2020 Deepfake Detection Challenge [Ferrer, 2020], for example, 2116 teams competed for one million dollars (USD) in prizes. Teams were provided 23,654 real videos and 104,500 deep-fake videos created from the provided real videos. The top performing learning-based detector achieved a detection accuracy of only 65% on a set of 4000 holdout videos, half of which were real and half of which were deep fakes (i.e., chance performance is 50%). These results reveal that fully automatic detection of deep fakes in the wild remains a challenging problem.

On the other hand, the advantage of artifact-based techniques is they can exploit inconsistencies that are difficult to circumvent or launder. The disadvantage is these techniques are typically narrowly applicable to a subset of deep-fake videos, and often require human annotation as part of the process.

The advantage of identity-based techniques is they are also resilient to adversarial and laundering attacks and are typically applicable to many different forms of deep fakes. The disadvantage of these approaches is an identity-specific model must be constructed for each individual, typically from hours of authentic video footage. This may be practical when it comes to, for example, protecting a few world leaders from deep fakes – for which hours of video can typically be found online – but is otherwise impractical. The other disadvantage is that the learned mannerisms are somewhat context dependent: when a world leader is giving a public address, for example, she may be more formal than when she is giving an unscripted interview, and so the specific mannerisms may not generalize across different contexts.

Because we are focused here on protecting one world leader – Ukrainian President Zelenskyy – and because we can easily acquire hours of video of Zelenskyy, we contend that an identity-based approach is the most sensible and robust approach. We start with the identity-based technique of [Agarwal et al., 2019], leveraging distinct patterns of facial and head movements, to distinguish Zelenskyy from an imposter or deep fake. We then augment this identity-based model with new gestural features capturing how a speaker uses their hands when speaking.

After reviewing the facial mannerisms portion of the model and describing the new gestural mannerisms portion, we evaluate the efficacy of our model in distinguishing Zelenskyy from deep-fake Zelenskyy and a range of other identities.

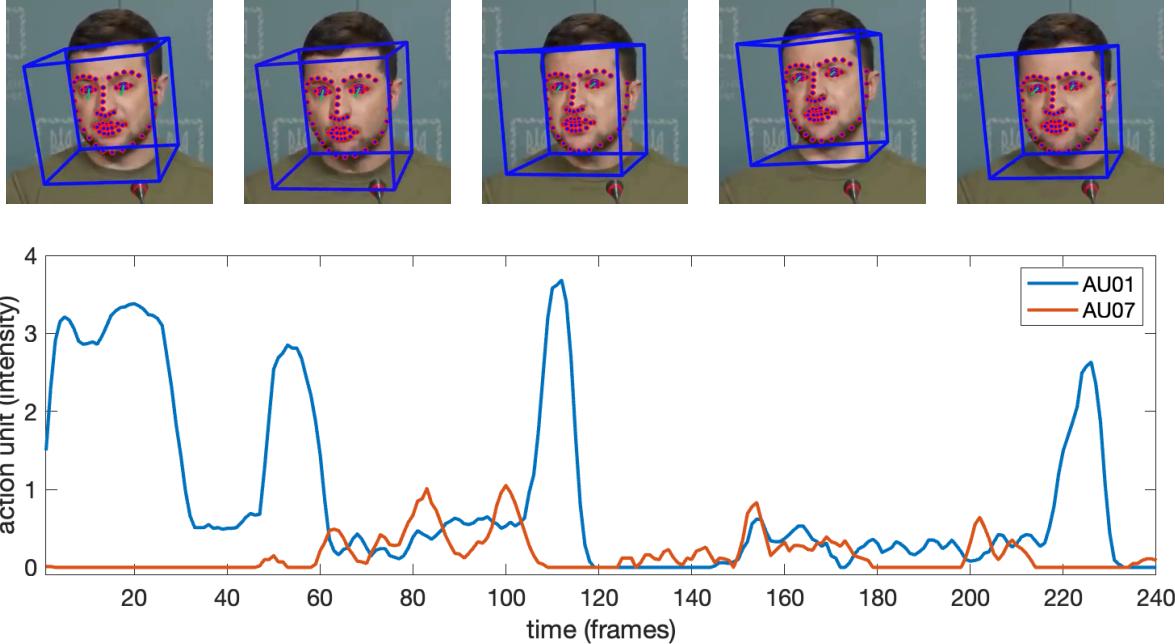


Figure 3: Shown above are five equally-spaced and cropped frames from a 10-second video clip annotated with the estimated facial landmarks (red markers) and head pose (blue box). Shown below are two of the 16 action units as a function of time (inner brow raiser [AU01] and lid tightener [AU07]).

2 Methods

2.1 Data Set

We downloaded 506 minutes of video of Zelenskyy from YouTube and the official website of the office of the Ukrainian president¹ in four different contexts: (a) public address (91 min); (b) press briefing (207 min); (c) bunker (47 min); and (d) armchair (161 min). Shown in Figure 2 are representative examples from each of these settings.

Portions of each video with large camera motions (e.g., zoom, translation, cross-fade) were automatically detected and removed from the data set. In particular, the inter-frame difference was computed between each successive pair of video frames. Assuming each video depicts a speaker in the center of the frame, a camera motion was detected if the absolute difference on the left and right margin (defined as 10% of the frame width) was above a specified threshold.

A total of 57 minutes of interview-style videos of seven world leaders (Jacinda Ardern, Joe Biden, Kamala Harris, Boris Johnson, Wladimir Klitschko, Angela Merkel, and Vladimir Putin) were used as decoys (i.e., not Zelenskyy). Our deep-fake detection is designed to distinguish Zelenskyy's behavioral and gestural mannerisms from imposters driving the creation of a deep fake, and so these decoy videos – regardless of the identities – serve as proxies for deep fakes. An additional 50 minutes of video across 27 distinct individuals taken from the FaceForensics++ [Rössler et al., 2019] dataset were used as additional decoys. In addition to these proxies, three commissioned lip-sync deep fakes (2 min) created by the team at Colossyan², and one in-the-wild deep fake (1 min) were added to this decoy dataset (Figure 1).

2.2 Facial Mannerisms

The identity-based forensic technique of [Agarwal et al., 2019] is based on the observation that individuals have distinct speaking styles in terms of facial expressions and head movements. Former President Obama, for example, tends to tilt his head upwards when he smiles, and downwards when he frowns.

Starting with a single video as input, the OpenFace2 toolkit [Baltrusaitis et al., 2018] extracts facial landmark positions, facial action units, head pose, and eye gaze on a per-frame basis. Facial muscle movement and expression are encoded using facial action units (AU) [Ekman and Friesen, 1976]. The OpenFace2 toolkit provides – on a per-frame basis – the

¹<https://www.president.gov.ua/en/videos/videos-archive>

²<https://www.colossyan.com>

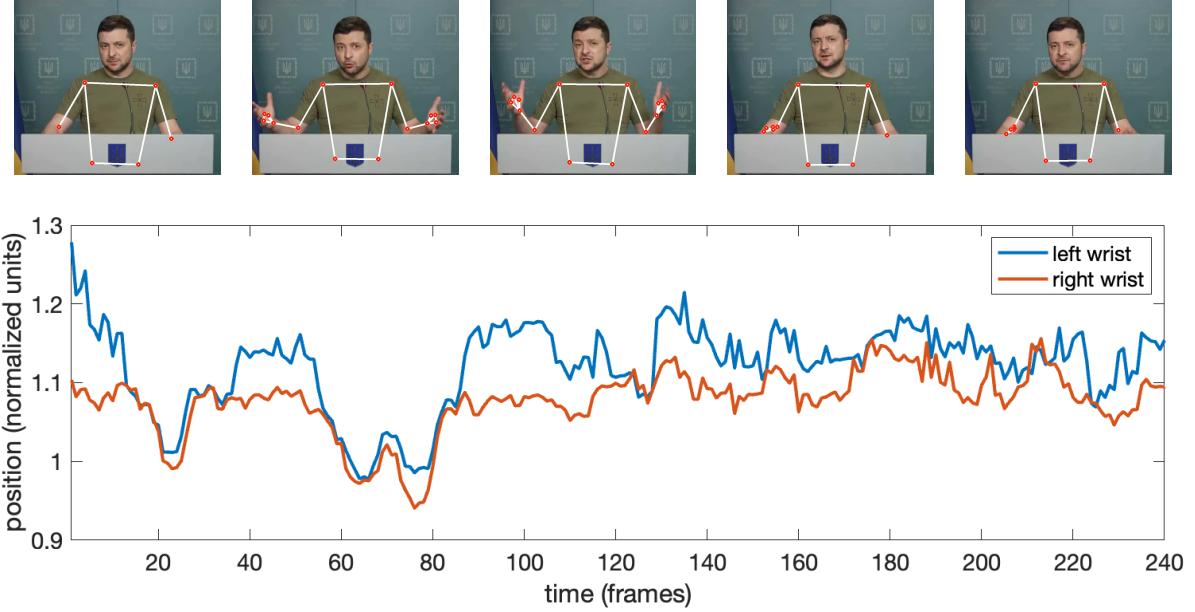


Figure 4: Shown above are five equally-spaced frames from a 10-second video clip annotated with the estimated gestural tracking. Shown below are two of the 12 gestural features corresponding to the vertical position of the left and right wrist (the spatial position of the wrists are reported in normalized units relative to a body-centric action plane).

strength of 17 different AUs: inner brow raiser (AU01), outer brow raiser (AU02), brow lowerer (AU04), upper lid raiser (AU05), cheek raiser (AU06), lid tightener (AU07), nose wrinkle (AU09), upper lip raiser (AU10), lip corner puller (AU12), dimpler (AU14), lip corner depressor (AU15), chin raiser (AU17), lip stretcher (AU20), lip tightener (AU23), lip part (AU25), jaw drop (AU26), and eye blink (AU45).

The forensic facial model incorporates 16 AUs (AU45, eye blink, was found not to be distinctive and therefore eliminated from consideration) and four additional features: (1) head rotation about the x-axis; (2) head rotation about the z-axis (as with AU45, head rotation about the y-axis was found not to be distinctive); (3) the horizontal distance between the corners of the mouth (mouth_h); and (4) the vertical distance between the lower and upper lip (mouth_v), yielding a total of 20 facial-mannerism features. Shown in Figure 3 are several frames of the facial and head tracking and two representative examples of the measured action units across a 10-second clip.

These features are combined with gestural mannerisms, described next, to form a person-specific behavioral model.

2.3 Gestural Mannerisms

Across cultures and languages, hand gestures provide additional information not always captured by a speaker’s words alone [Church et al., 2017]. In addition to distinct gestural patterns found across age [Özer D. et al., 2017], sex [Özçalışkan and Goldin-Meadow, 2010], and culture [Pika et al., 2006], recent work also finds that individuals exhibit distinct gestural patterns [Özer D and T., 2020]. We, therefore, hypothesize that hand gestures, in addition to the facial expressions and head movements described above, will improve our ability to identify an individual’s distinct speaking patterns.

Arm and hand position and movement are estimated in each input video frame using Blazepose [Bazarevsky et al., 2020] from the MediaPipe library [Lugaresi et al., 2019]. Because we are interested only in the upper body, we consider the image x -, y -coordinates corresponding to the shoulder, elbow, and wrist of both arms, Figure 4, yielding a total of 12 individual measurements. These upper-body coordinates, initially specified relative to the video-frame size, are normalized into a speaker-centric action plane [Boháček and Hrúz, 2022]. This action plane is a rectangular bounding box centered on the speaker’s chest with a width $8 \times$ and height $6 \times$ the measured head height [De Silva, 2008, Bauer, 2014]. In this normalized bounding box, the upper left-hand corner is $(0, 0)$ and the lower right-hand corner is $(1, 1)$. This normalization ensures that the tracked upper-body coordinates can be compared across different speaker locations and sizes. Shown in Figure 4 are several frames of the upper-body tracking and representative examples of the extracted gestural features across a 10-second clip.

Model	World	Real	Lip-Sync	In-The-Wild
	Leaders	FF++	Deep-Fake	Deep-Fake
facial	91.7	91.1	94.7	17.4
gestural	77.4	95.7	95.0	12.1
facial + gestural	100.0	100.0	97.1	94.9
DFDC	73.1	84.5	93.5	13.3
			Zelenskyy	Zelenskyy

Table 1: Classification accuracy (reported as percentages) for our behavioral model with facial, gestural, and these two features combined evaluated against seven different world leaders, 28 identities in the FaceForensics++ dataset and against both real and deep-fake versions of Zelenskyy. By comparison, our model significantly outperforms the best-performing DFDC model [Seferbekov, 2020].

Whereas the tracked x, y facial features are converted into a higher-level representation in the form of action units, we find that a similar approach with the hand gestures was less effective than simply considering the normalized x, y locations of the tracked shoulders, elbows, and wrists.

2.4 Behavioral Model

Correlations between all pairs of the 20 facial features and 12 gestural features are used to capture individualized mannerisms (e.g., head tilt and smiling/frowning). A total of ${}_{32}C_2 = (32 \times 31)/2 = 496$ correlations are extracted from overlapping 10-second video clips extracted from an input video in question.

Trained on authentic video of a person of interest, a novelty detection model in the form of a one-class, non-linear support vector machine (SVM) [Schölkopf et al., Pedregosa et al., 2011] is used to distinguish an individual from imposters and deep fakes. An advantage of this classifier is that it only requires examples of authentic videos.

The 506 minutes of Zelenskyy video is partitioned into overlapping (by 5 seconds) 10-second video clips, yielding a total of 157,752 clips. The 110 minutes of other identities in the World Leaders, FaceForensics++, and Deep-Fake Zelenskyy videos are similarly partitioned, yielding a total of 25,077 clips.

These clips are randomly partitioned into a 80/20 training/testing split. The SVM is trained on the 496 facial- and gestural-feature pairwise correlations. The SVM hyper-parameters, consisting of the Gaussian kernel width (γ) and outlier percentage (ν), are optimized by performing a grid search over these parameters across the training set. The trained classifier is then evaluated against the hold-out testing set. This entire process is repeated 100 times with randomized training/testing splits, from which we report average classification accuracy.

Three different classifiers are trained on facial features only, gestural features only, and facial and gestural features combined. The SVM classification threshold for the individual features is selected to yield a 95% training accuracy of correctly classifying real Zelenskyy clips. The classification threshold for the combined features is selected to yield a 99% training accuracy.

3 Results

Shown in Table 1 is the classification accuracy (averaged over 100 random training/testing data splits) of our behavioral model evaluated against the 10-second video clips of seven different world-leaders, 28 distinct identities in the FaceForensics++ dataset [Rössler et al., 2019], and real and deep-fake versions of Zelenskyy.

We find that the facial features and gestural features alone are insufficient to consistently detect deep-fake version of Zelenskyy (see the last two columns of Table 1). The combination of facial and gestural, however, yields significant improvements in detection accuracy. Because deep-fake techniques are – rightfully – focused on high-quality facial and audio synthesis, and because of the expected difficulty in synthesizing realistic hands and hand gestures, we posit that the combination of facial and gestural signals will prove reliable for at least a few years.

As compared to the best-performing DFDC model (last row of Table 1) [Seferbekov, 2020], our model achieves significantly higher classification across all non-Zelenskyy data sets. This comparison, however, is not entirely fair as our behavioral model is trained to detect deep-fake versions of just one identity, whereas the DFDC model is a generic deep-fake detector. On the other hand, our classifier operates on 10-second video clips whereas the DFDC model has the advantage of operating on the entire video. This comparison does, nevertheless, show the power of identity-specific models.

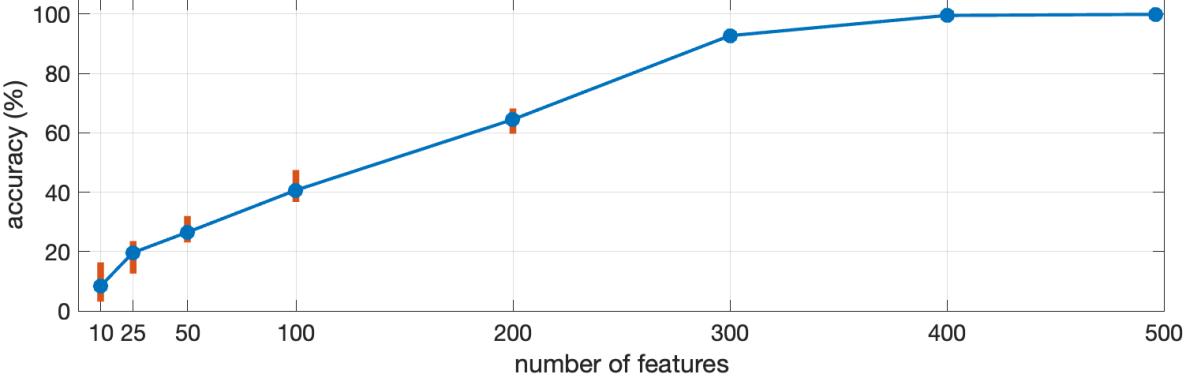


Figure 5: Each data point corresponds to the median (50% quantile) accuracy for classifiers trained on between 10 and 496 randomly selected facial and gestural features; the error bars correspond to the 25% and 75% quantile.

3.1 Ablation

To determine how many of the 496 pairwise facial and gestural correlations are needed to achieve the classification accuracy reported in Table 1, we trained a series of one-class SVMs on randomly selected subsets – ranging in size between 10 and 400 – of all facial and gestural features. Shown in Figure 5 is the median accuracy of classifying the identities in the world leaders and deep-fake Zelenskyy videos. In this figure, each data point corresponds to the median accuracy (50% quantile) from 25 independent and randomly selected features of each subset size; the error bars correspond to the 25% and 75% quantile.

With the full set of 496 facial and gestural features, detection accuracy is 99.88%. Detection accuracy grows relatively linearly between feature subsets of size 10 and 300 plateauing at 99.52% with 400 features. Here we see that a significant fraction of the facial and gestural features are, collectively, rich and informative.

To determine which specific facial and gestural features are most discriminative, we next trained 500 classifiers on random feature subsets of size 10. The discriminatory power of each feature is computed from the average accuracy of each classifier to which a feature contributed. Across all 500 classifiers, the detection accuracy on the world leaders and deep-fake Zelenskyy data sets ranges from 44.4% to 4.3%. The top 20 most discriminative correlation features and respective classifier accuracy are:

Feature 1	Feature 2	Classifier Accuracy (%)	
head-pose-Rx	\Leftrightarrow	right-elbow-y	44.4
head-pose-Rx	\Leftrightarrow	right-wrist-y	36.9
head-pose-Rx	\Leftrightarrow	left-elbow-y	33.7
left-elbow-x	\Leftrightarrow	left-shoulder-x	32.8
head-pose-Rx	\Leftrightarrow	left-shoulder-y	32.5
head-pose-Rx	\Leftrightarrow	lip-vertical	29.3
left-elbow-x	\Leftrightarrow	right-elbow-y	27.4
right-elbow-y	\Leftrightarrow	right-shoulder-y	27.4
right-elbow-x	\Leftrightarrow	right-shoulder-x	27.3
head-pose-Rx	\Leftrightarrow	left-wrist-y	26.9
AU14	\Leftrightarrow	AU17	26.2
AU06	\Leftrightarrow	right-elbow-y	25.6
mouth _v	\Leftrightarrow	AU14	25.5
AU12	\Leftrightarrow	AU15	25.4
AU12	\Leftrightarrow	AU14	25.0
right-elbow-y	\Leftrightarrow	left-shoulder-y	25.0
pose-Rz	\Leftrightarrow	right-shoulder-x	24.4
mouth _v	\Leftrightarrow	AU15	23.7
pose-Rx	\Leftrightarrow	right-shoulder-y	23.0
AU06	\Leftrightarrow	AU14	22.7

where * -Rx and * -Rz correspond to 3-D head rotations, * -x and * -y correspond to the horizontal and vertical image position, and AU * corresponds to specific facial action units (see Section 2.2). Here we see that the correlation between

head rotation and hand gestural features are the most discriminative, highlighting the importance of the addition of gestural features to the original facial-based model. For President Zelenskyy, in particular, head rotation (as in nodding affirmatively) is highly correlated to his hand movements.

As compared to the median accuracy of 8.4% across random features of subset size 10 (Figure 5), these top-ranked features achieve accuracies between three and five times higher. A single classifier trained on the top 10 and 20 features, however, only yields a prediction accuracy on the world leaders and deep-fake Zelenskyy data sets of 59.2% and 63.4%, providing further evidence that a full set of facial and gestural features are necessary to achieve a high classification accuracy.

4 Discussion

Although the term deep fakes first splashed on the screen in 2017, the precursor to what we now call deep fakes dates back two decades. In the seminal video-rewrite work [Bregler et al. \[1997\]](#), a video of a person speaking is automatically modified to yield a video of them saying things not found in the original footage. The resulting video quality and resolution were generally lower than today’s deep-fake videos, but the results were nevertheless impressive. Some 25 years later, deep neural networks, GANs, massive data sets, and unlimited compute cycles have led to increasingly more realistic and sophisticated deep-fake videos.

While the democratization of access to techniques for manipulating and synthesizing videos has led to interesting and entertaining applications, they have also given rise to complex ethical and legal question [Chesney and Citron \[2019\]](#). In the fog of war, in particular, deep fakes pose a significant threat to our ability to understand and respond to rapidly evolving events.

While our approach to protecting a single individual – Ukrainian President Zelenskyy – does not address the broader issue of deep fakes, it does bring some level of digital protection to the arguably most important Ukrainian voice at this time of war.

Acknowledgement

We are grateful to Zoltan Kovacs, Muhammad Shahzaib Aslam, and the rest of the Colossyan team for creating the Zelenskyy lip-sync deep fakes.

References

- Bobby Allyn. Deepfake video of Zelenskyy could be ‘tip of the iceberg’ in info war, experts warn. <https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia>, 2022.
- Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1), 2022.
- Cristian Canton Ferrer. Deepfake detection challenge results: An open initiative to advance AI. <https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai>, 2020.
- Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection. In *International Conference on Computer Vision and Pattern Recognition*, 2017.
- Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: A compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, 2018.
- Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-ray for more general face forgery detection. arXiv:1912.13458, 2019.
- Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In *International Workshop on Information Forensics and Security*, pages 1–7, 2018.
- Shruti Agarwal and Hany Farid. Detecting deep-fake videos from aural and oral dynamics. In *CVPR Workshop on Media Forensics*, pages 981–989, 2021.
- Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *CVPR Workshop on Media Forensics*, pages 660–661, 2020a.

- Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR Workshop on Media Forensics*, volume 1, 2019.
- Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *International Workshop on Information Forensics and Security*, pages 1–6, 2020b.
- Shruti Agarwal, Liwen Hu, Evonne Ng, Trevor Darrell, Hao Li, and Anna Rohrbach. Watch those words: Video falsification detection using word-conditioned facial motion. arXiv:2112.10936, 2021.
- Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. ID-reveal: Identity-aware deepfake video detection. In *International Conference on Computer Vision*, pages 15108–15117, 2021.
- Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *CVPR Workshop on Media Forensics*, pages 658–659, 2020.
- Mauro Barni, Matthew C Stamm, and Benedetta Tondi. Adversarial multimedia forensics: Overview and challenges ahead. In *European Signal Processing Conference*, pages 962–966, 2018.
- Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision and Pattern Recognition*, 2019.
- Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 59–66, 2018.
- Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1(1):56–75, 1976.
- R. Breckinridge Church, Marha W. Alibali, and Spencer D. Kelly. *Why Gesture?: How the hands function in speaking, thinking and communicating*, volume 7. John Benjamins Publishing Company, 2017.
- Özer D., Tansan M., Özer E. E., Malykhina K., Chatterjee A., and Göksun T. The effects of gesture restriction on spatial language in young and elderly adults. In *38th Annual Conference of the Cognitive Science Society*, pages 1471—1476, 2017.
- Şeyda Özçalışkan and Susan Goldin-Meadow. Sex differences in language first appear in gesture. *Developmental Science*, 13(5):752–760, 2010.
- Simone Pika, Elena Nicoladis, and Paula F Marentette. A cross-cultural study on the use of gestures: Evidence for cross-linguistic transfer? *Bilingualism: Language and Cognition*, 9(3):319–327, 2006.
- Özer D and Göksun T. Gesture use and processing: A review on individual differences in cognitive resources. *Frontiers in Psychology*, 2020.
- Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. arXiv:2006.10204, 2020.
- Camillo Lugaressi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. arXiv:1906.08172, 2019.
- Matyáš Boháček and Marek Hruž. Sign pose-based transformer for word-level sign language recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 182–191, 2022.
- Liyanage De Silva. Audiovisual sensing of human movements for home-care and security in a smart environment. *International Journal On Smart Sensing and Intelligent Systems*, 1, 2008.
- Anastasia Bauer. *The Use of Signing Space in a Shared Sign Language of Australia*. De Gruyter, 1 edition, 2014.
- Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Selim Seferbekov. DFDC deepfake challenge solution. https://github.com/selimsef/dfdc_deepfake_challenge, 2020.
- Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *24th Annual Conference on Computer Graphics and Interactive Techniques*, pages 353–360, 1997.
- Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107:1753, 2019.