Introduction to Recognizing Deep Fakes

Mark Sherman
Director, Cybersecurity Foundations, CERT
April 4, 2022
AI4Cyber Conference

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Copyright 2022 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® and CERT® are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM22-0235

Outline

Introduction

Making Deep Fakes

Machine Recognition of Deep Fakes

Visual Recognition of Deep Fakes

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Carnegie Mellon Leads an Ecosystem of Innovation for Cybersecurity

CMU Campus – Global Research University

- Global research university known for its world-class, interdisciplinary programs in computer science, machine learning/artificial intelligence, engineering, business, arts, policy, and science
- Ranked #1 Artificial Intelligence, #1 Cybersecurity, #1 Software Engineering, #1 Computer Engineering, #2 Computer Science, Programming Languages, #3 Computer Systems, Data Analytics, Theory
  (*U.S. News and World Report*)
- 1,442 total faculty and 130 research centers
- CyLab, CMU's security and privacy research institute, brings together experts from all schools across the university

CMU Software Engineering Institute (SEI)

- Founded in 1984 by the DoD as a Federally-Funded Research and Development Center (FFRDC) focused on software engineering
- Leader in software engineering, cybersecurity, and artificial intelligence research
- Established CERT in 1988
- About $145M annual funding (~$23M DoD Line)
- Critical to the DoD ability to acquire, develop, operate, and sustain software systems that are innovative, affordable, trustworthy, and enduring *(CMU SEI Sponsoring Agreement)*

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

CERT Division

Founded on a unique combination of experiential understanding of DoD missions, the cyber warfighter, the operational domain, and constantly changing technology

Adapts the best science to impact operational missions, increase the trustworthiness of technology, and develop cyber talent

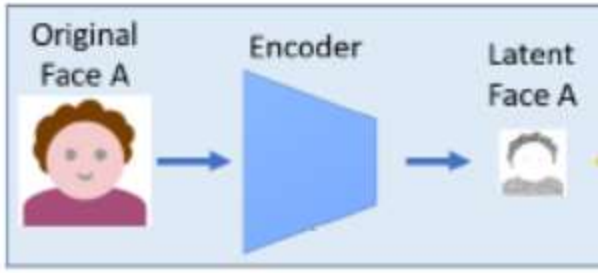
Partners with DoD, non-DoD agencies, and the private sector enable CERT to maintain technical depth, attract top talent, amplify DoD financial investment, reduce the risk to DoD missions, and scale the research

Strengthens the resilience of critical national functions, increases the cybersecurity and resilience of DoD systems and Defense Industrial Base, and develops the cyber capacity of allies and partners

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Creating Deep Fakes using Autoencoders



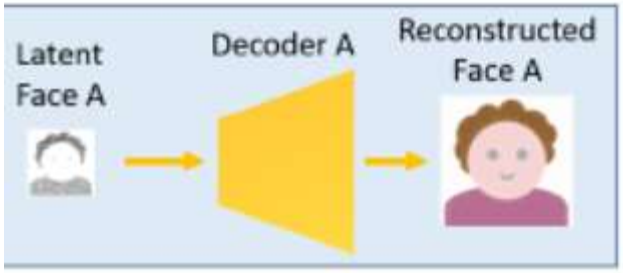
- First neural net “learns” features of Face A (master)
- i.e., output layer represents features of Face A

T. Nguyen, C. Nguyen, D. Nguyen, D. Nguyen, and S. Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv:1909.11573v2 [cs.CV] 28 Jul 2020, https://dev.arxiv.org/abs/1909.11573v2

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Creating Deep Fakes using Autoencoders



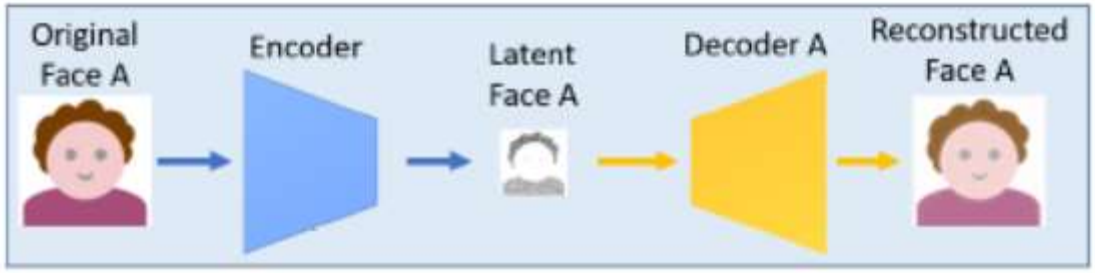
- Second neural net “learns” to reconstruct Face A
- i.e., output layer represents pixels of Face A

T. Nguyen, C. Nguyen, D. Nguyen, D. Nguyen, and S. Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv:1909.11573v2 [cs.CV] 28 Jul 2020, https://dev.arxiv.org/abs/1909.11573v2

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Creating Deep Fakes using Autoencoders



- First neural net "learns" features of Face A (master)
- i.e., output layer represents features of Face A

- Second neural net "learns" to reconstruct Face A
- i.e., output layer represents pixels of Face A

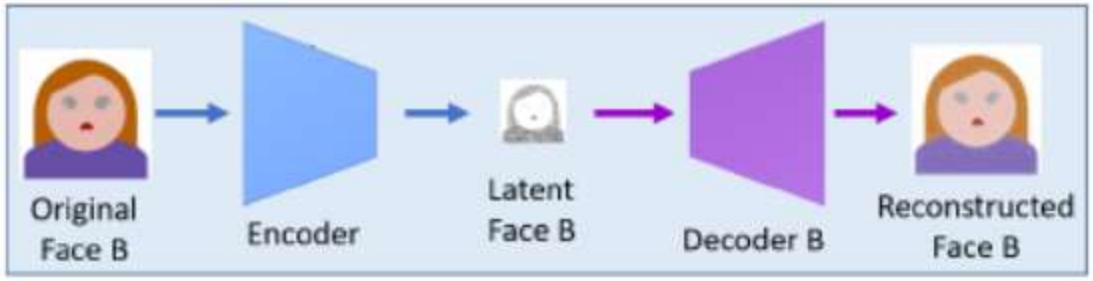
Together, the two neural nets can encode and reconstruct a "master"

T. Nguyen, C. Nguyen, D. Nguyen, D. Nguyen, and S. Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv:1909.11573v2 [cs.CV] 28 Jul 2020, https://dev.arxiv.org/abs/1909.11573v2

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Creating Deep Fakes using Autoencoders

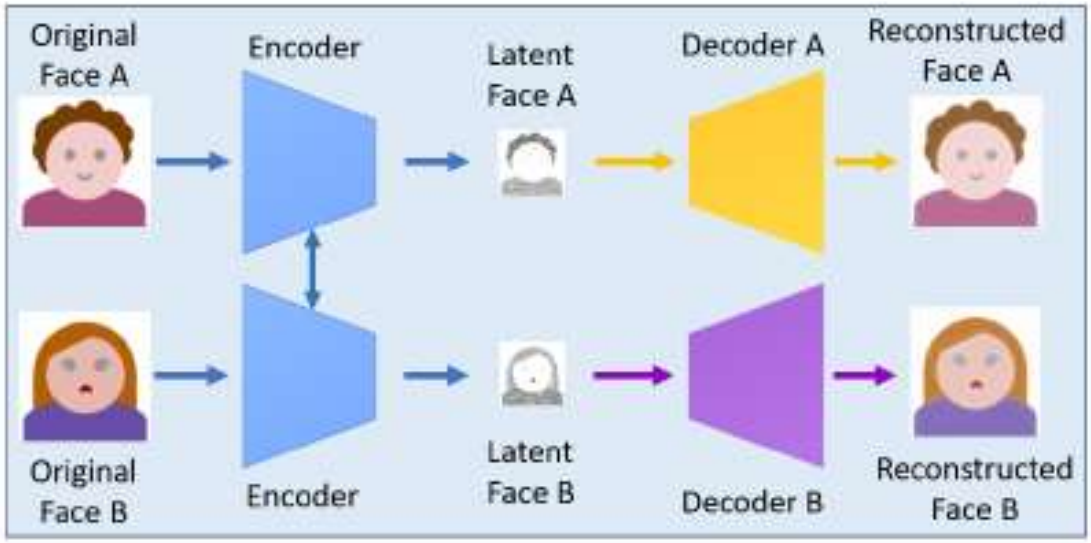


By the same process, two neural nets can encode and reconstruct a “target”

T. Nguyen, C. Nguyen, D. Nguyen, D. Nguyen, and S. Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv:1909.11573v2 [cs.CV] 28 Jul 2020, https://dev.arxiv.org/abs/1909.11573v2

© 2022 Carnegie Mellon University
[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Creating Deep Fakes using Autoencoders



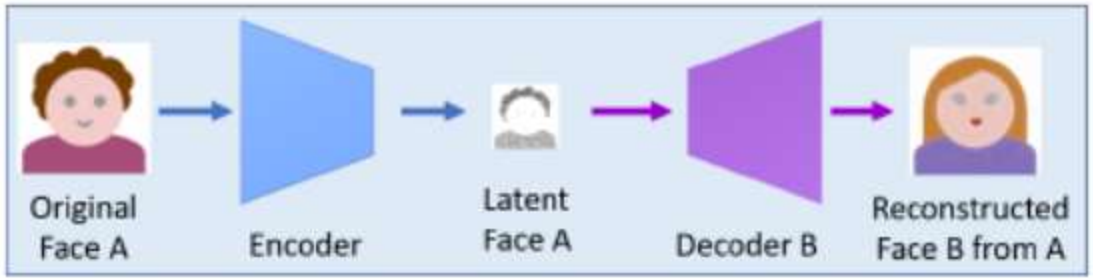
In application, the two “encoders” are the same, i.e., trained on both master and target to produce common features

T. Nguyen, C. Nguyen, D. Nguyen, D. Nguyen, and S. Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv:1909.11573v2 [cs.CV] 28 Jul 2020, https://dev.arxiv.org/abs/1909.11573v2

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Creating Deep Fakes using Autoencoders



Using the master to generate features which feed the decoder specific to the target creates the fake

T. Nguyen, C. Nguyen, D. Nguyen, D. Nguyen, and S. Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv:1909.11573v2 [cs.CV] 28 Jul 2020, https://dev.arxiv.org/abs/1909.11573v2

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Visual Detection of Deep Fakes

Illustration source:
https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

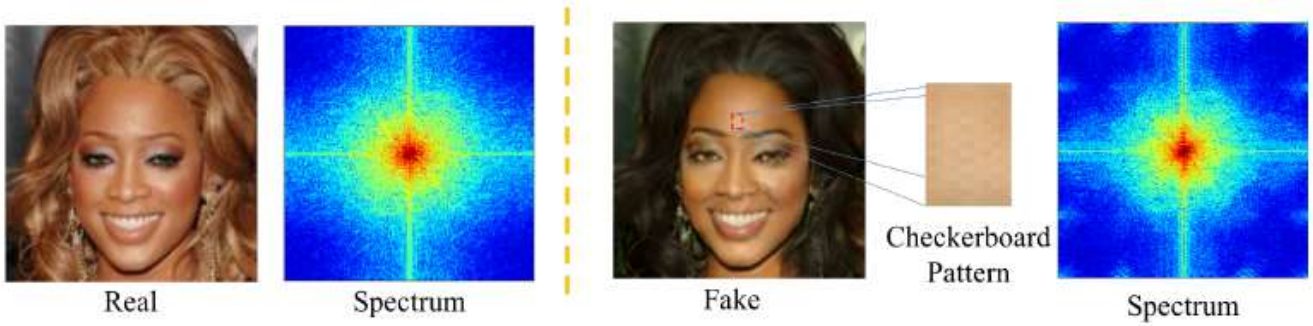
Some artifacts can be detected visually

- Mismatched skin tone at borders, especially for face swaps
- Inconsistent light sources and shadows
- “Lip reading” (visemes) and sounds (phonemes) aka lip sync
- Discontinuities in video
- Inconsistent eye blinking
- Mismatched jewelry, e.g., earrings
- Selectively indistinct features

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Machine Detection of Deep Fakes – GAN Artifacts



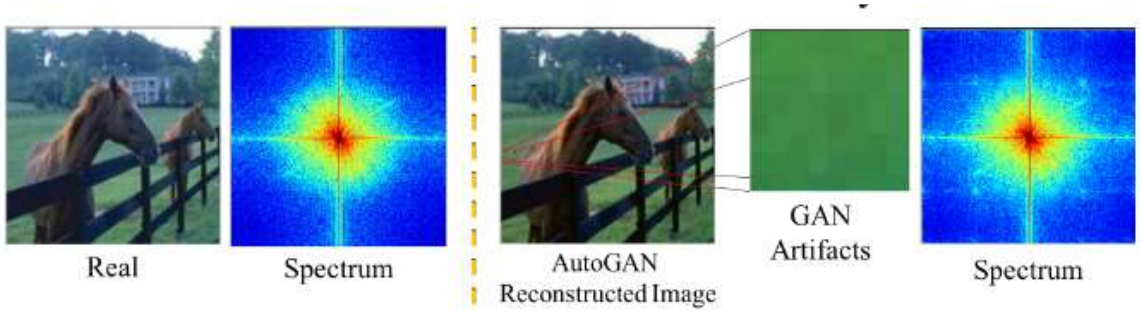
Source: X. Zhang, S. Karaman, and S. Chang, "Detecting and Simulating Artifacts in GAN Fake Images (Extended Version)," 15 Oct 2019, https://arxiv.org/abs/1907.06515

- Mathematical fingerprints in fakes
  - GANs leave unique markers, such as frequency spectra in generated images (checkboard)
    - Up-sampler detection
    - Interpolation detection

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Machine Detection of Deep Fakes – GAN Artifacts



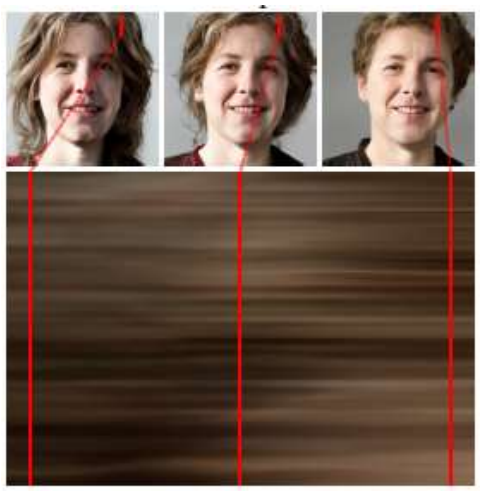
Source: X. Zhang, S. Karaman, and S. Chang, "Detecting and Simulating Artifacts in GAN Fake Images (Extended Version)," 15 Oct 2019, https://arxiv.org/abs/1907.06515

- Mathematical fingerprints from original
  - AutoGANs can create versions of original image
  - Difference “fingerprint” can be used to train detector

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Machine Detection of Deep Fakes – Physiological Inconsistency



Source: T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-Free Generative Adversarial Networks," 35th Conference on Neural Information Processing Systems, 2021, https://arxiv.org/abs/2106.12423

Gross checks

- Mismatch between facial movements and speech
- Mismatch between facial movements and respiration
- Inconsistent detected heart beats

Scale differences

- Gross anatomical features determine facial layout
- Details are interpolated unnaturally

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Many proposed ways to detect fakes

Table 3. Summary of Deepfake Detection Models

			Type	Modality			Content			Method			Eval. Dataset							Performance*		
			Reenactment	Replacement	Image	Video	Audio	Feature	Body Part	Face Image	Model	Indicates Affected Area	Input Resolution	DeepfakeTIMIT [86]	DFFD [149]	FaceForensics [130]	FaceForensics++ [131]	FFW [82]	Celeb-DF [101]	Other Deepfake DB	Custom	ACC	EER	AUC
Classic ML	[187]	2017	•		•	•				•	SVM-RBF		250x250								•	92.9		
	[4]	2017	•		•	•				•	SVM		*							•			18.2	
	[178]	2018	•		•	•				•	SVM		*							•				0.97
	[86]	2018	•		•					•	SVM		128x128	•									3.33	
	[42]	2019			•	•				•	SVM, Kmeans...		1024x1024								•	100		
	[8]	2019	•		•	•				•	SVM		*		•								13.33	
	[6]	2019	•		•					•	SVM		*								•			0.98
Deep Learning	[111]	2018	•		•	•				•	CNN		256x256								•	99.4		
	[97]	2018	•		•			•	•		LSTM-CNN		224x224								•			0.99
	[119]	2018	•		•	•				•	Capsule-CNN		128x128			•						99.3		
	[17]	2018			•	•				•	ED-GAN		128x128								•	92		
	[39]	2018	•		•					•	CNN		1024x1024							•				0.81
	[63]	2018	•		•		•			•	CNN-LSTM		299x299								•	97.1		
	[106]	2018	•		•					•	CNN		256256								•	94.4		
	[33]	2018	•		•	•				•	CNN AE		256x256			•					•	90.5		
	[3]	2018	•		•					•	CNN		256x256								•			0.99
	[132]	2019	•		•					•	CNN-LSTM		224x224				•				•	96.9		
	[118]	2019	•		•	•				•	CNN-DE	•	256x256			•	•				•	92.8	8.18	
	[38]	2019			•	•				•	CNN		-								•	98.5		
	[41]	2019	•		•	•				•	CNN AE GAN	•	256x256				•				•	99.2		
	[149]	2019	•		•	•				•	CNN+Attention	•	299x299		•								3.11	0.99
	[98]	2019	•		•	•				•	CNN		128x128							•	•			0.99
	[101]	2019			•	•				•	CNN		*						•					0.64
	[52]	2019	•		•	•				•	CNN+HMN		224x224			•	•	•				99.4		
	[92]	2019	•		•	•				•	FCN		256x256				•					98.1		
	[177]	2019	•		•	•				•	CNN		128x128								•	94.7		
	[161]	2019	•		•	•				•	CNN		224x224							•		86.4		
	[153]	2019	•		•					•	CNN		1024x1024								•			94
	[30]	2019	•		•		•			•	CNN		128x128			•					•	96		
	[99]	2019	•		•	•				•	CNN	•	224x224	•									93.2	
	[11]	2019	•		•					•	CNN		224x224				•					81.6		
	[?]	2019	•			•	•	•	•		LSTM		*								•		22	
	[47]	2019	•			•	•	•	•		LSTM-DNN		*								•		16.4	
	[25]	2019	•			•				•	CNN		256x256								•	97		
	[180]	2019	•		•	•			•	•	CNN		128x128			•		•			•	99.6	0.53	
	[166]	2019	•		•	•				•	SVM+VGGnet		224x224				•					85		
	[94]	2019	•		•	•			•		CNN	•	64x64		•		•		•	•				99.2
	[95]	2020	○		•	•				•	HRNet-FCN	•	64x64		•		•		•				20.86	0.86
	[96]	2020	•		•	•				•	PP-CNN		-			•			•					0.92
	[123]	2020	•		•	•				•	ED-CNN		299x299				•		•	•				0.99
	[108]	2020	•		•		•			•	ED-LSTM		224x224				•		•					
	[167]	2020	•		•	•				•	CNN ResNet		224x224				•				•	Avrg.	Prec.=	0.93
	[64]	2020	•		•	•				•	AREN-CNN		128x128				•				•	98.52		
	[110]	2020	•		•		•	•	•	•	ED-CNN		*	•	•					•				0.92
	[5]	2020	•		•		•	•		•	CNN		128x128							•		89.6		
	[10]	2020	•		•		•			•	LSTM		256x256				•					94.29		
	[69]	2020	•		•	•				•	Siamese CNN		64x64								•	TPR=0.91		
	[129]	2020	•		•					•	Ensemble		224x224				•					99.65		1.00
	[36]	2020	•		•	•				•	*		112x112						•			98.26		99.73
	[81]	2020	•		•	•				•	OC-VAE		100x100				•					TPR=0.89		
	[51]	2020	•		•					•	ABC-ResNet		224x224								•		?	
Statistics & Steganalysis	[85]	2018	•		•					•	PRNU		1280x720								•	*TPR=1*	*FPR=*	*0.03*
	[150]	2019	•		•	•				•	Statistics	•	-								•			
	[107]	2019	•	•	•					•	PRNU		*								•	90.3		

*Only the best reported performance, averaged over the test datasets, is displayed to capture the 'best-case' scenario.

Mirksy & Lee, "The Creation and Detection of Deepfakes: A Survey", Sept 13, 2020, https://arxiv.org/pdf/2004.11138.pdf

Blending (spatial)

- Edge detectors, quality measures, frequency analysis

Environmental (spatial)

- Face warping, lighting, varying fidelity

Forensics (spatial)

- Network fingerprints, camera/sensor noise, inconsistent head poses

Behavior (temporal)

- Mannerism anomalies, perceived emotion

Physiology (temporal)

- Heart rate, blood volume, eye blinking

Synchronization (temporal)

- Audio/video matching, mouth shapes

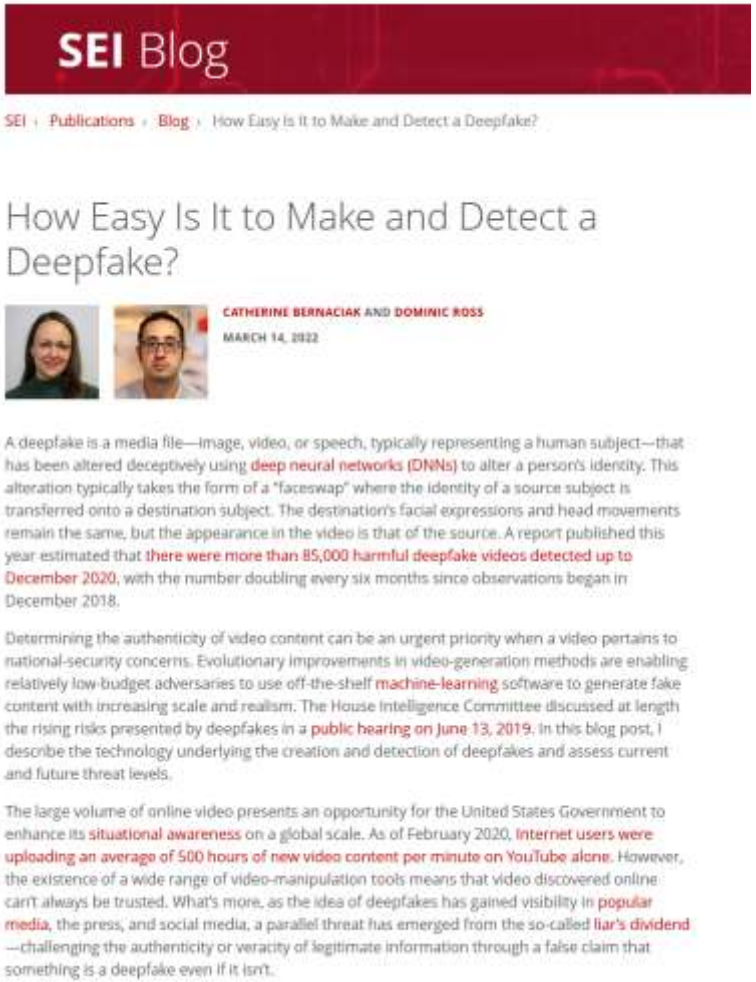
Coherence (temporal)

- Flickers, jitter, frame prediction

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Learning more about deep fakes



SEI Blog

SEI › Publications › Blog › How Easy Is It to Make and Detect a Deepfake?

How Easy Is It to Make and Detect a Deepfake?

CATHERINE BERNACIAK AND DOMINIC ROSS

MARCH 14, 2022

A deepfake is a media file—image, video, or speech, typically representing a human subject—that has been altered deceptively using deep neural networks (DNNs) to alter a person's identity. This alteration typically takes the form of a "faceswap" where the identity of a source subject is transferred onto a destination subject. The destination's facial expressions and head movements remain the same, but the appearance in the video is that of the source. A report published this year estimated that there were more than 85,000 harmful deepfake videos detected up to December 2020, with the number doubling every six months since observations began in December 2018.

Determining the authenticity of video content can be an urgent priority when a video pertains to national-security concerns. Evolutionary improvements in video-generation methods are enabling relatively low-budget adversaries to use off-the-shelf machine-learning software to generate fake content with increasing scale and realism. The House Intelligence Committee discussed at length the rising risks presented by deepfakes in a public hearing on June 13, 2019. In this blog post, I describe the technology underlying the creation and detection of deepfakes and assess current and future threat levels.

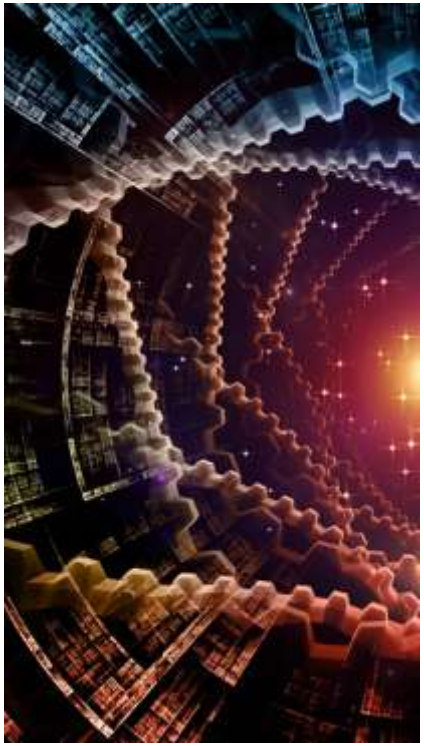
The large volume of online video presents an opportunity for the United States Government to enhance its situational awareness on a global scale. As of February 2020, Internet users were uploading an average of 500 hours of new video content per minute on YouTube alone. However, the existence of a wide range of video-manipulation tools means that video discovered online can't always be trusted. What's more, as the idea of deepfakes has gained visibility in popular media, the press, and social media, a parallel threat has emerged from the so-called liar's dividend —challenging the authenticity or veracity of legitimate information through a false claim that something is a deepfake even if it isn't.

https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Ways to Engage with Us



- **Download software and tools**
- **Explore research and capabilities**
- **Participate in education offerings**
- **Attend an event**
- **Search the digital library**
- **Read the SEI Year in Review**
- **Collaborate with the SEI on a new project**

Software Engineering Institute
Carnegie Mellon University
4500 Fifth Avenue
Pittsburgh, PA 15213-3890
412-268-5800 - Phone
888-201-4479 - Toll-Free
412-268-5758 - Fax
info@sei.cmu.edu - Email
www.sei.cmu.edu - Web

© 2022 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.