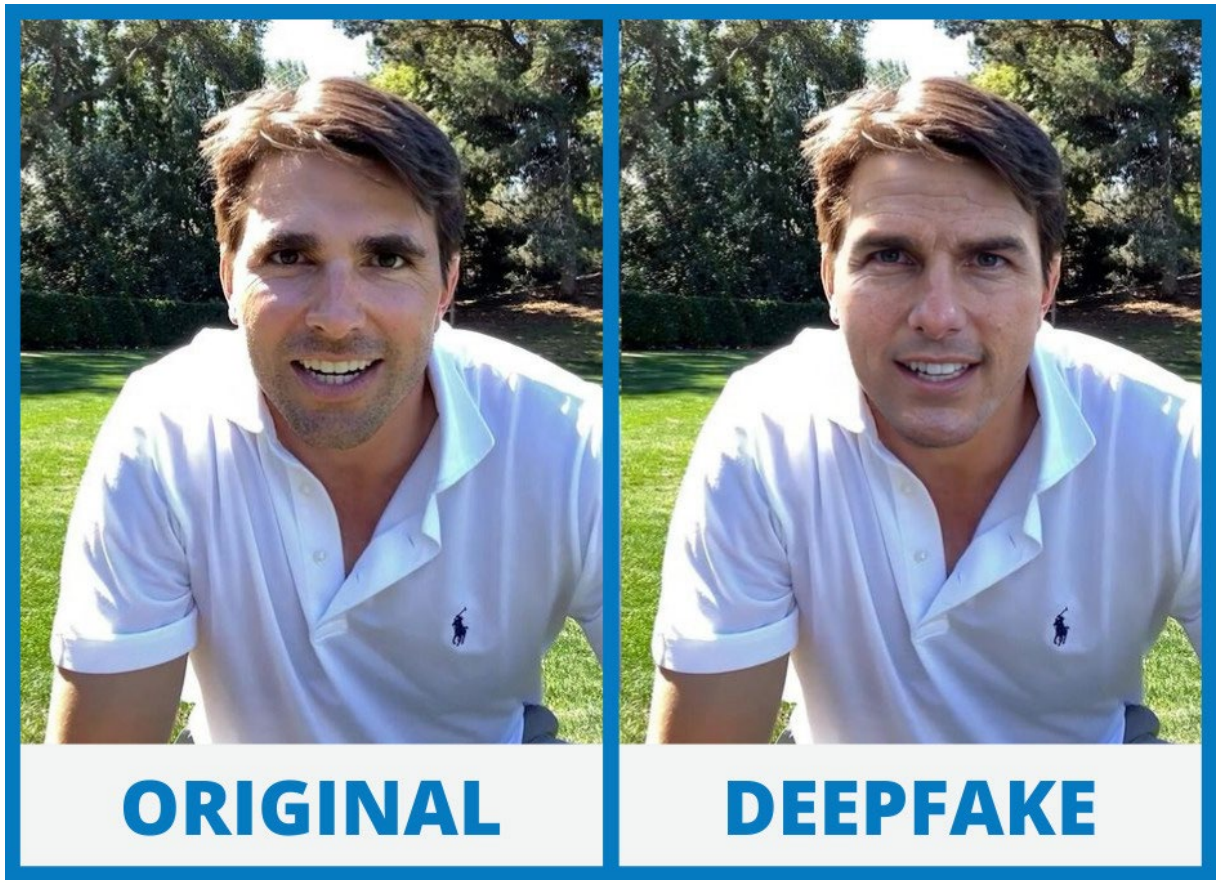
What You See Is Not What You Know: Deepfake Image Manipulation

Cathryn Allen, graduate student at KSU
Dr. Bryson Payne, Professor of CS/Cyber, UNG
Dr. Tamirat Abegaz, Professor of CS/Cyber, UNG
Dr. Chuck Robertson, Professor of Psychological Sciences, UNG

UNG
UNIVERSITY of NORTH GEORGIA™

What are deepfakes?

- Manipulated media
- Contain misinformation



Why do we care about deepfakes?

- Progressive expansion and dependency on computers (McPeak, 2021).
- Can become blind to false news
- Need awareness and knowledge

Previous Findings

- Deepfakes are on the rise
- Misinformation spreads easily and is believable (Anderson, 2020).
- More information can help stop it from spreading (Pennycook and Rand, 2020).
- People are more drawn to believe visuals than textuals (Frenda et al., 2013).
- Online is a perfect breeding ground (Anderson, 2020).

Hypothesis 1 and 2: Familiarity

- If the subjects of the deepfakes are someone that the viewer is familiar with, the viewer will be more capable of determining if a video is a deepfake.
- If the subjects of the deepfakes are someone that the viewer is unfamiliar with, the viewer will be less capable of determining if a video is a deepfake.

Hypothesis 3 and 4: Quantity

- If the viewers are presented with two videos, one original and one deepfake, they will be more capable of determining which video is the deepfake.
- If the viewer is only presented one video at a time, whether it being an original video or a deepfake video, the viewer will be less capable of determining if the video is a deepfake.

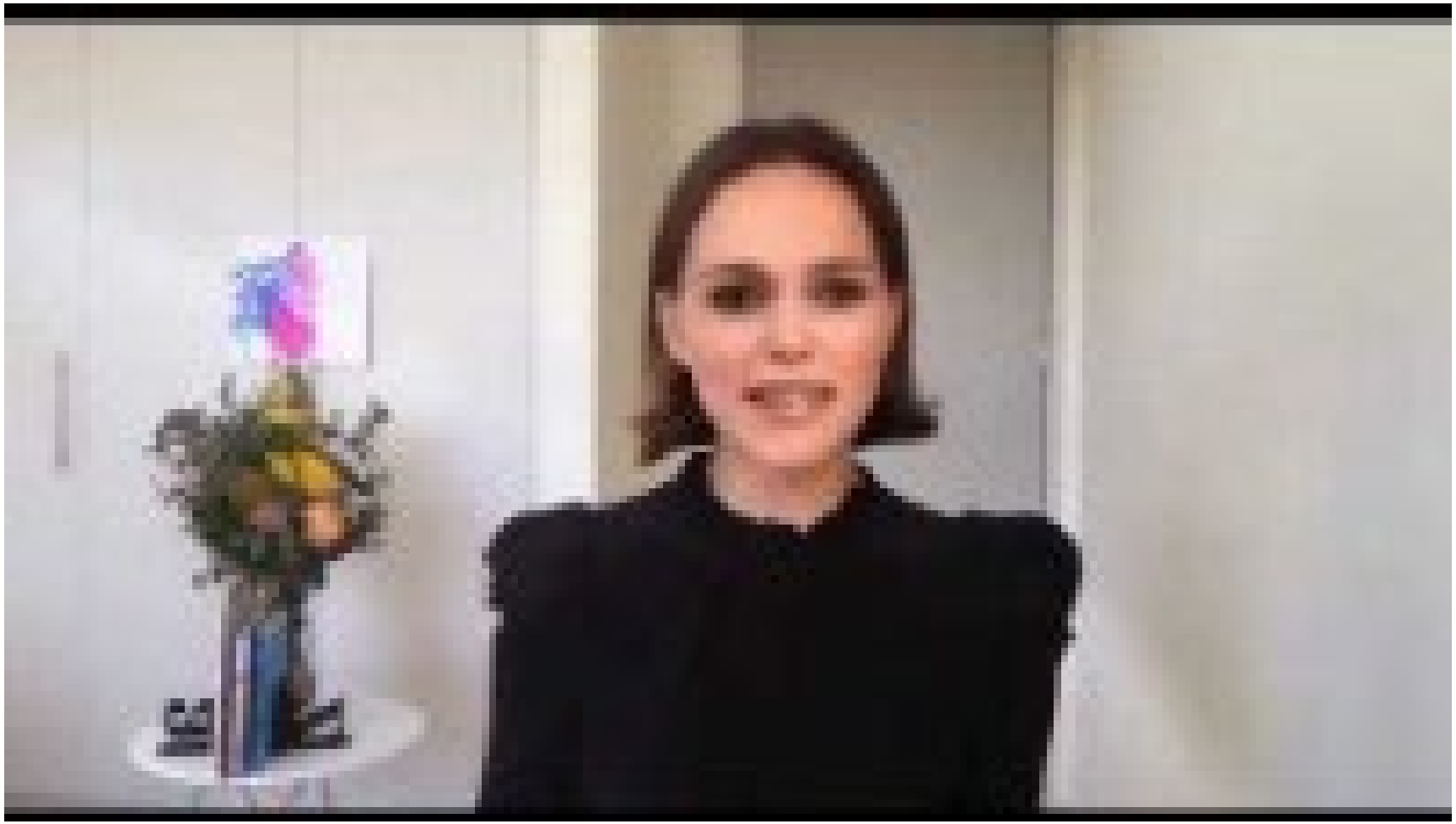
Survey

- Qualtrics
- Gain an understanding of deepfakes
- 4 question sets
  - One Video Unfamiliar
  - One Video Familiar
  - Two Video Unfamiliar
  - Two Video Familiar
- Follow up Questions after the video(s)
  - Will walk through examples

Deepfakes for the survey

- Video collection
- DeepFaceLab (Chervoniy et al.).
- Pipeline workflow: multiple ways to approach a solution
  - Extraction
  - Training
  - Conversion
- Can take a more automated or manual approach

One Video Familiar



UNG | UNIVERSITY of NORTH GEORGIA™

Two Video Unfamiliar

UNG | UNIVERSITY of NORTH GEORGIA™

Results

- Any data that was not fully complete was thrown out
- If a participant did not answer the correct celebrity, their data was ignored for the question when familiarity mattered
- Only 52/154 (33.77%) people could determine all 4 videos correctly
  - Means 66.23% of people were fooled by one or more video

Review of Hypothesis 1&2: Familiarity

	Unfamiliar Accuracy	Familiar Accuracy
One video at a time	43.9%	72.7%
Two videos side-by-side	82.2%	95.3%

- T-test was performed on accuracy between familiar and unfamiliar using an alpha of .05
  - The found t-value was 2.209, which is greater than the needed 1.96
  - Found to be a 5% significance level for the difference in familiarity

Review of Hypothesis 3&4: Quantity

	Unfamiliar Accuracy	Familiar Accuracy
One video at a time	43.9%	72.7%
Two videos side-by-side	82.2%	95.3%

- T-test was performed on accuracy between one video and two videos using an alpha of .05
  - The found t-value was 1.32, which is less than the needed 1.96
  - Found that quantity is not statistically significant

Conclusions

- Familiarity with subjects does play a role in how well someone can determine if a video is a deepfake.
- Quantity of videos has not shown an effect on how well someone can determine a deepfake.
- People are not that good at determining deepfakes vs original videos
  - Only 33.77% (52/154) could get all 4 correct
- Addressing Deepfakes could be better achieved by helping a viewer become more familiar with the subject of the deepfake

Countering Deepfakes: Case Study

- Volodymyr Zelensky – in March 2022 during the first six weeks of the Russian invasion of Ukraine, a deepfake video of a false Zelensky appeared to urge Ukrainians to surrender to the Russian forces. Within minutes of a television station's mention of the video, President Zelensky posted a Facebook video discrediting the deepfake video and denying the video's message.
- Based on our research, it is possible that Zelensky's use of live video of himself (increasing the viewer's familiarity with him as the subject) was as important as the content of the repudiation itself.

Refuting Deepfake Disinformation: Guidance for Organizations

- Organizations, governments, and individuals seeking to contain or counter deepfake deception will need to consider both factors in Zelensky's case study for their operational planning:
  1. a swift, near-real-time response, and
  2. creating more familiarity through additional, preferably **live video footage** of the target of the deepfake responding to and refuting the disinformation personally.

Future Work

- The authors intend to pursue additional research in countering deepfake videos:
  - determining whether disinformation in the form of a deepfake video of an individual could be addressed by providing additional, unaltered videos of the person to better familiarize the viewer with the target individual depicted in the deepfake.
- Providing this same “training” to computer models to aid in the automated detection of deepfakes is also an area of interest for future work.