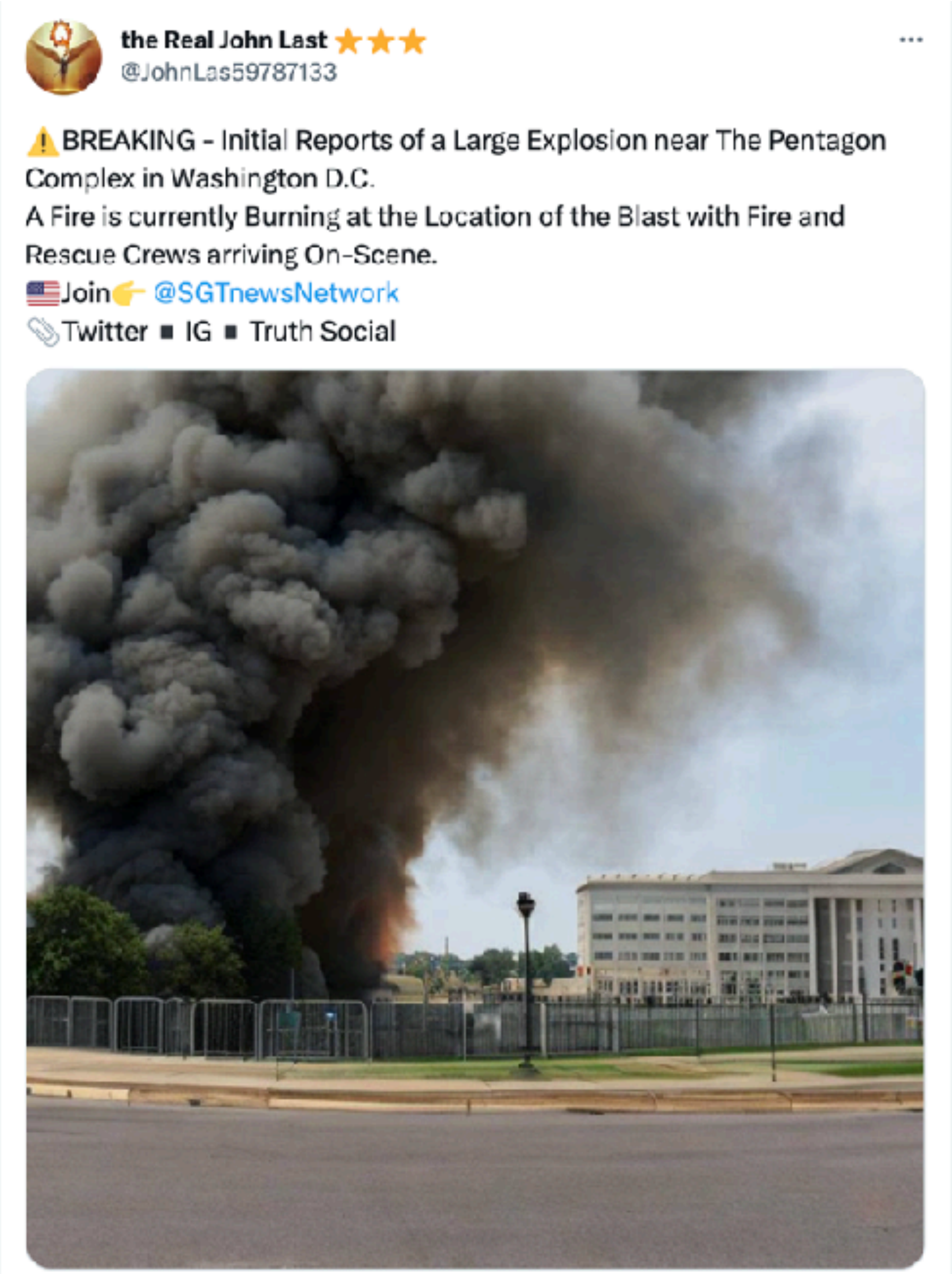
Combatting with DeepFakes

New Trends and Thoughts

Siwei Lyu, Ph.D.

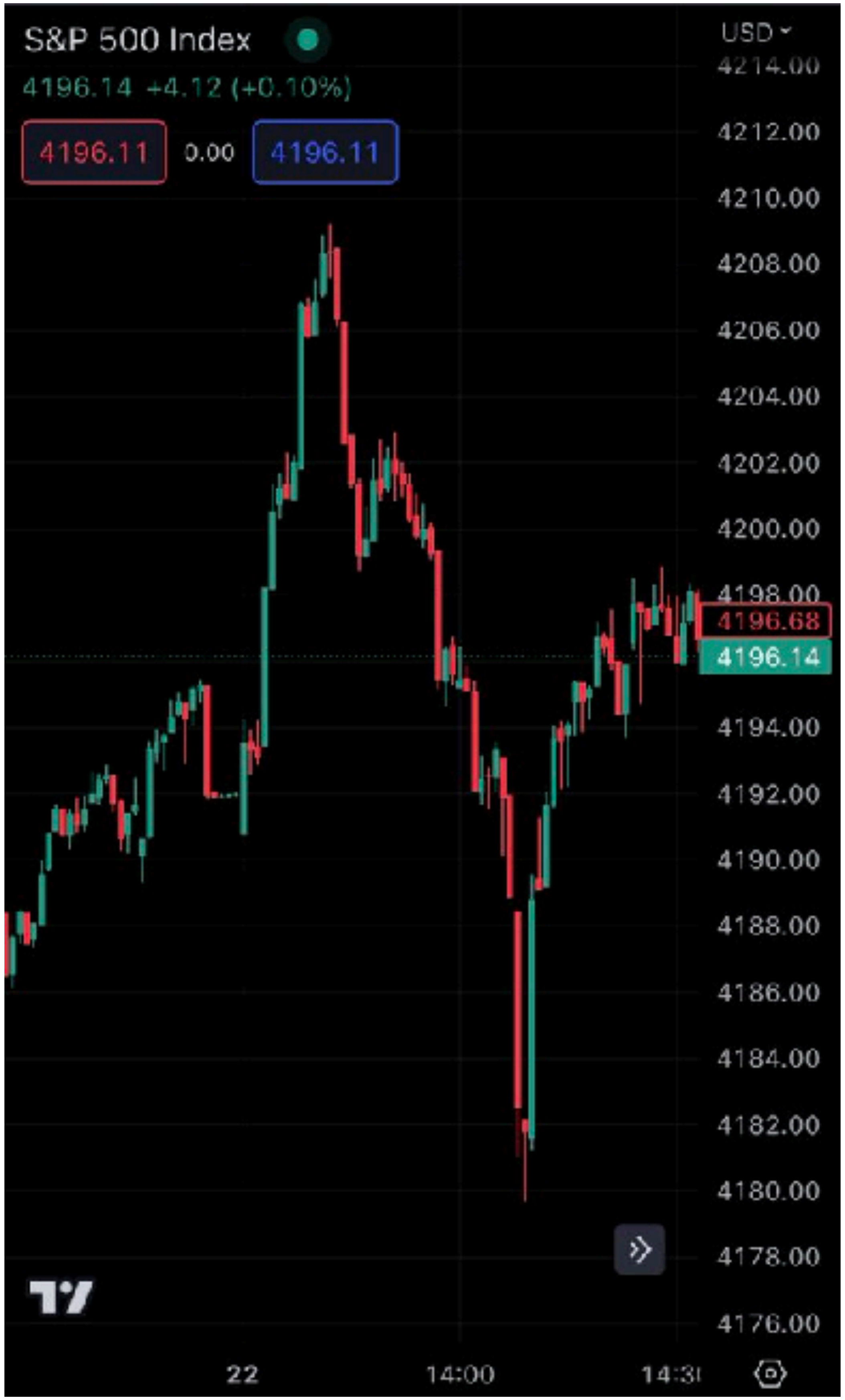
- Department of Computer Science and Engineering,
- Center for Information Integrity,
- Media Forensics Lab,
- University at Buffalo, State University of New York

May 21st, 2023

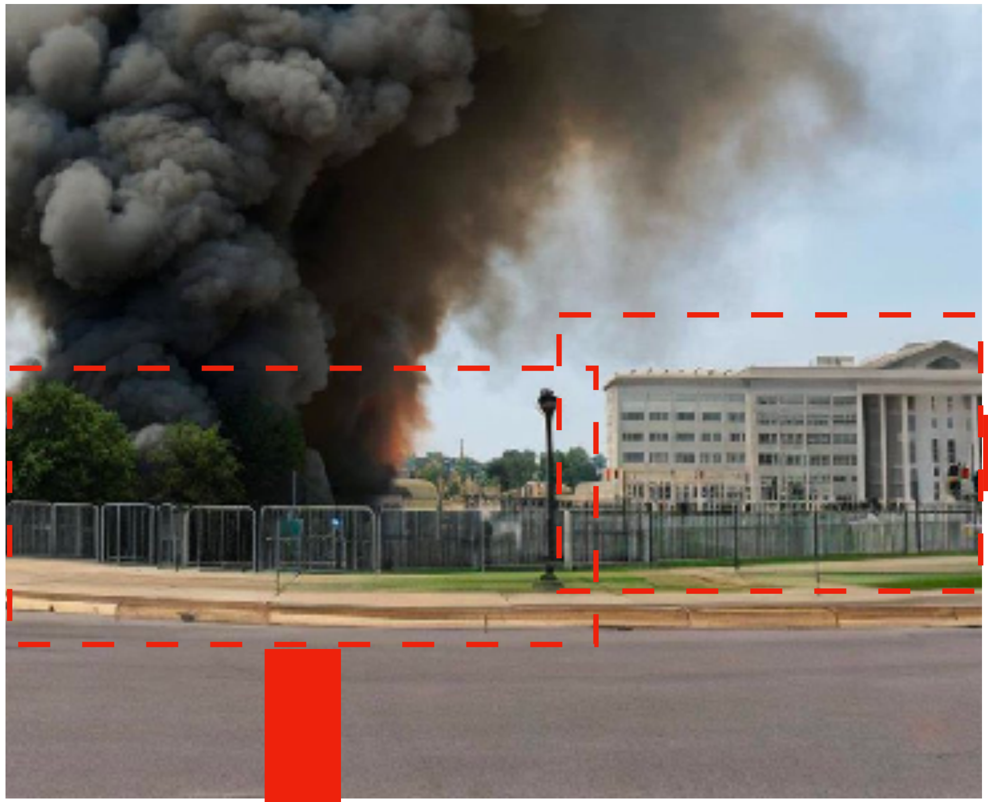


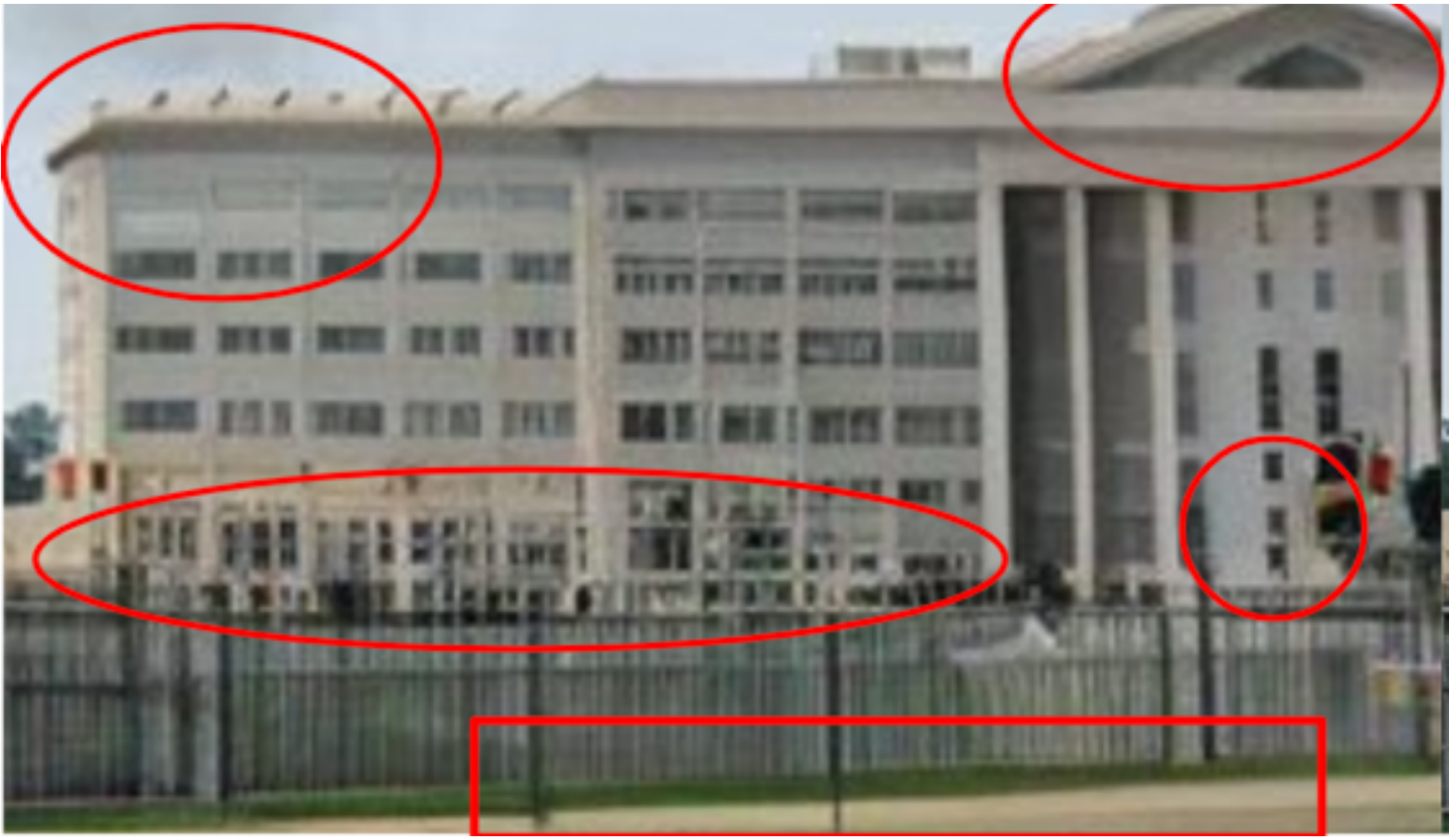
the Real John Last ★★★
@JohnLas59787133

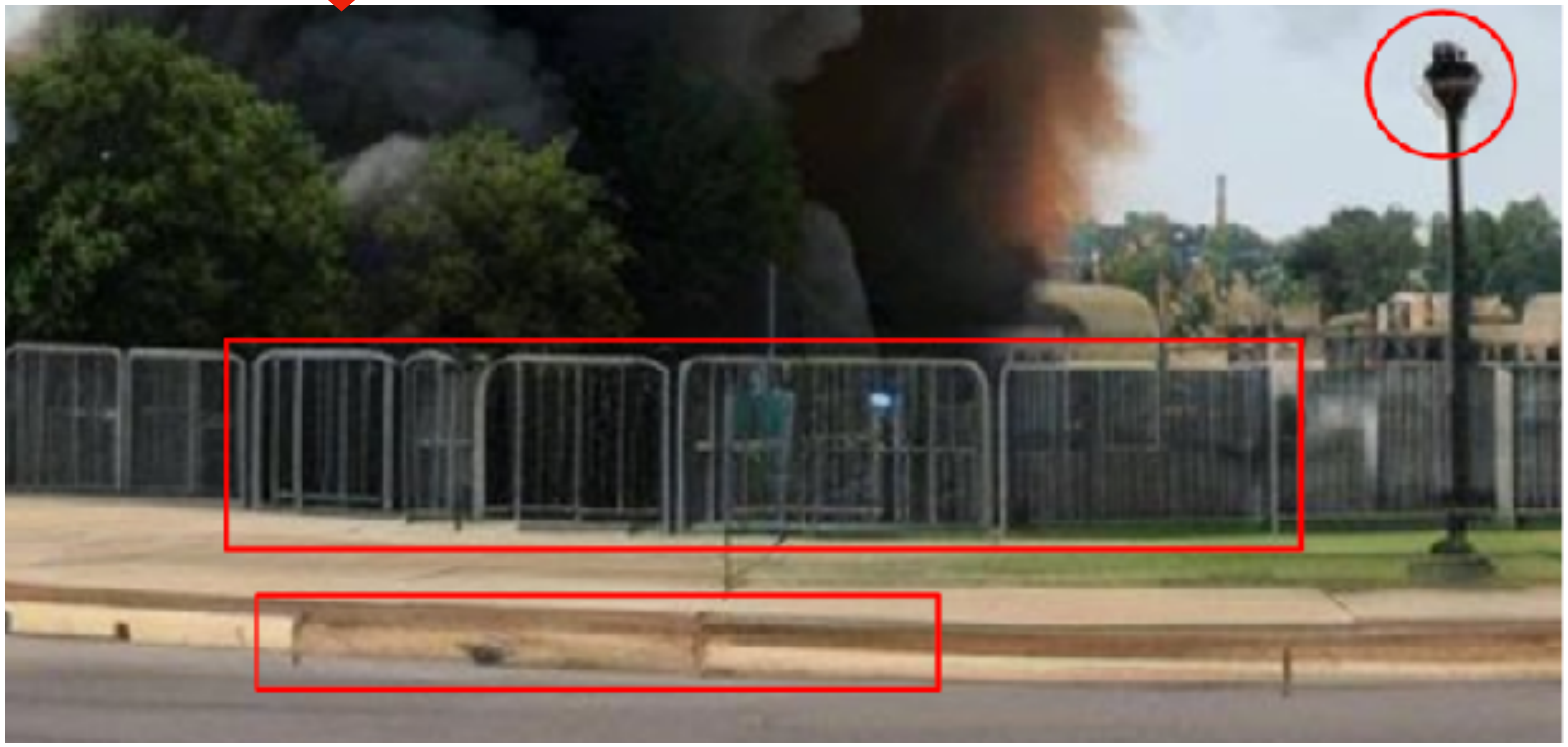
⚠️BREAKING - Initial Reports of a Large Explosion near The Pentagon Complex in Washington D.C.
A Fire is currently Burning at the Location of the Blast with Fire and Rescue Crews arriving On-Scene.
🇺🇸Join👉 @SGTnewsNetwork
📎Twitter ▪ IG ▪ Truth Social



May 21st, 2023

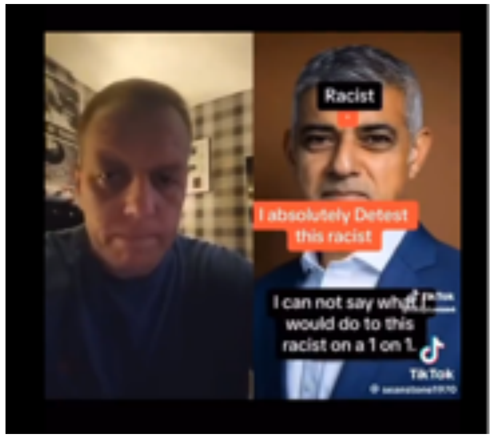






AI Generated

DeepFakes on the rise

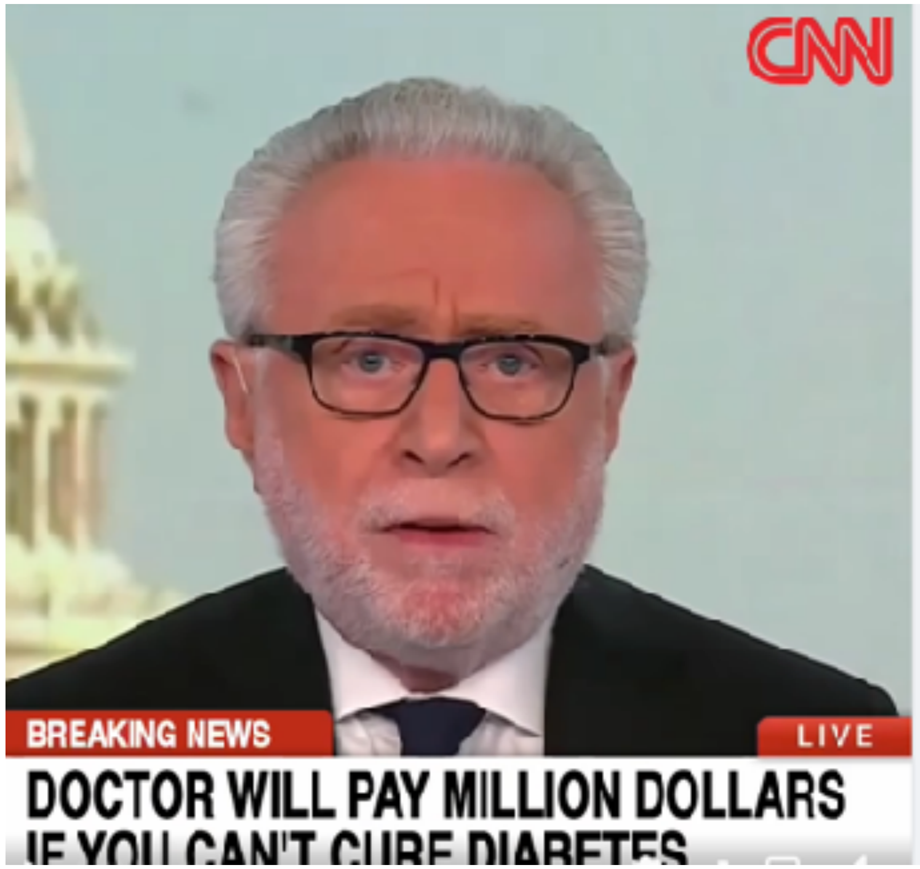

Racist
I absolutely Detest this racist
I can not say what I would do to this racist on a 1 on 1.
TikTok

Twitter 11/13/2023

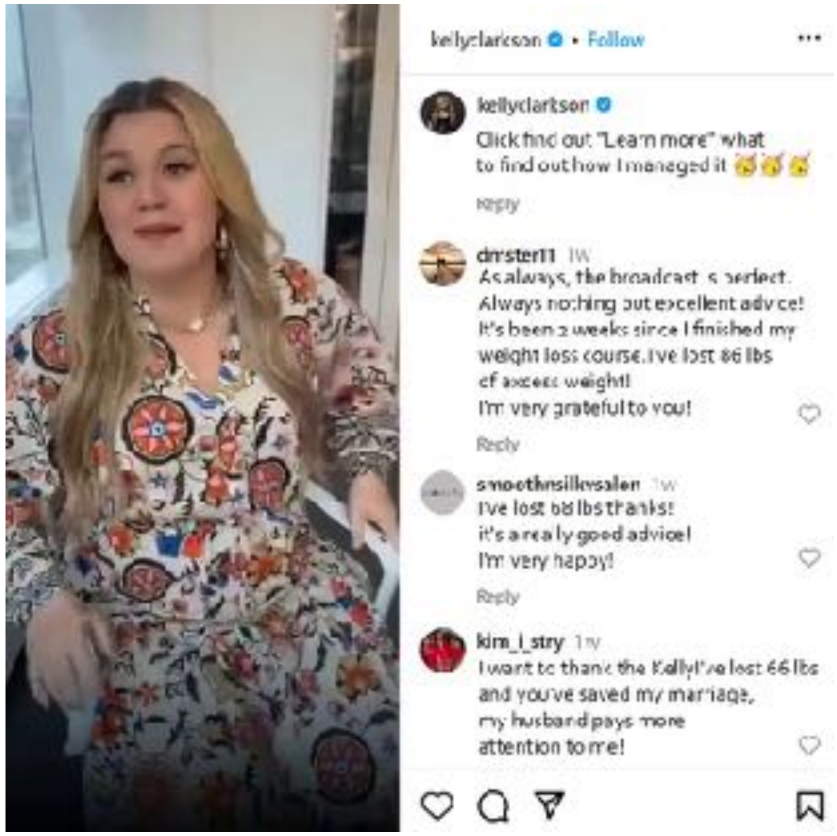
Twitter 11/06/2023

Twitter 10/30/2023

Twitter 10/20/2023



Facebook 11/06/2023



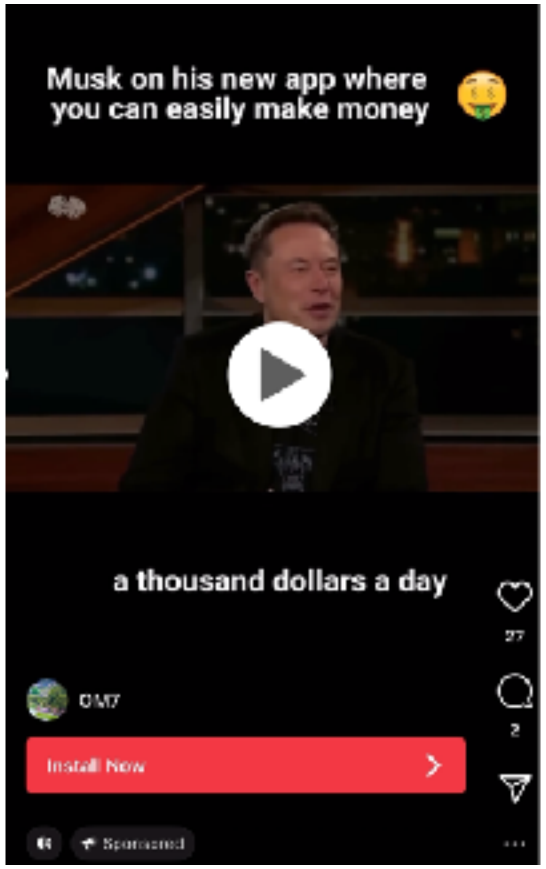
kellyclarkson
Click find out "Learn more" what to find out how I managed it 😿😿😿
Reply

drrster11 1W
As always, the broadcast is perfect. Always nothing but excellent advice! It's been 2 weeks since I finished my weight loss course. I've lost 86 lbs of excess weight!
I'm very grateful to you!
Reply

smoothnsilkysalon 1W
I've lost 68 lbs thanks!
it's a really good advice!
I'm very happy!
Reply

kim_i_stry 1W
I want to thank the Kelly! I've lost 66 lbs and you've saved my marriage, my husband pays more attention to me!

Facebook 10/27/2023



Twitter 10/22/2023

Rise of AIGC

Science & engineering

VAE Transformer Autoregression

GAN Normalized Flow Diffusion

Tools & services

Midjourney ChatGPT ElevenLabs

Data source & spreading channels

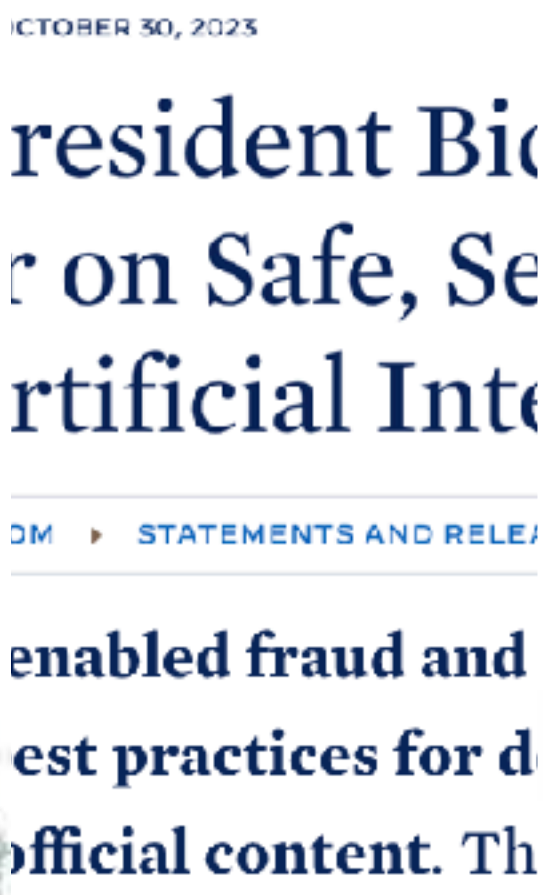
Computation power

aws Microsoft Azure

White House Executive Order on AI

Administration Priorities The Record

Gen AI



)CTOBER 30, 2023

resident Bi(
r on Safe, Se
rtificial Int(

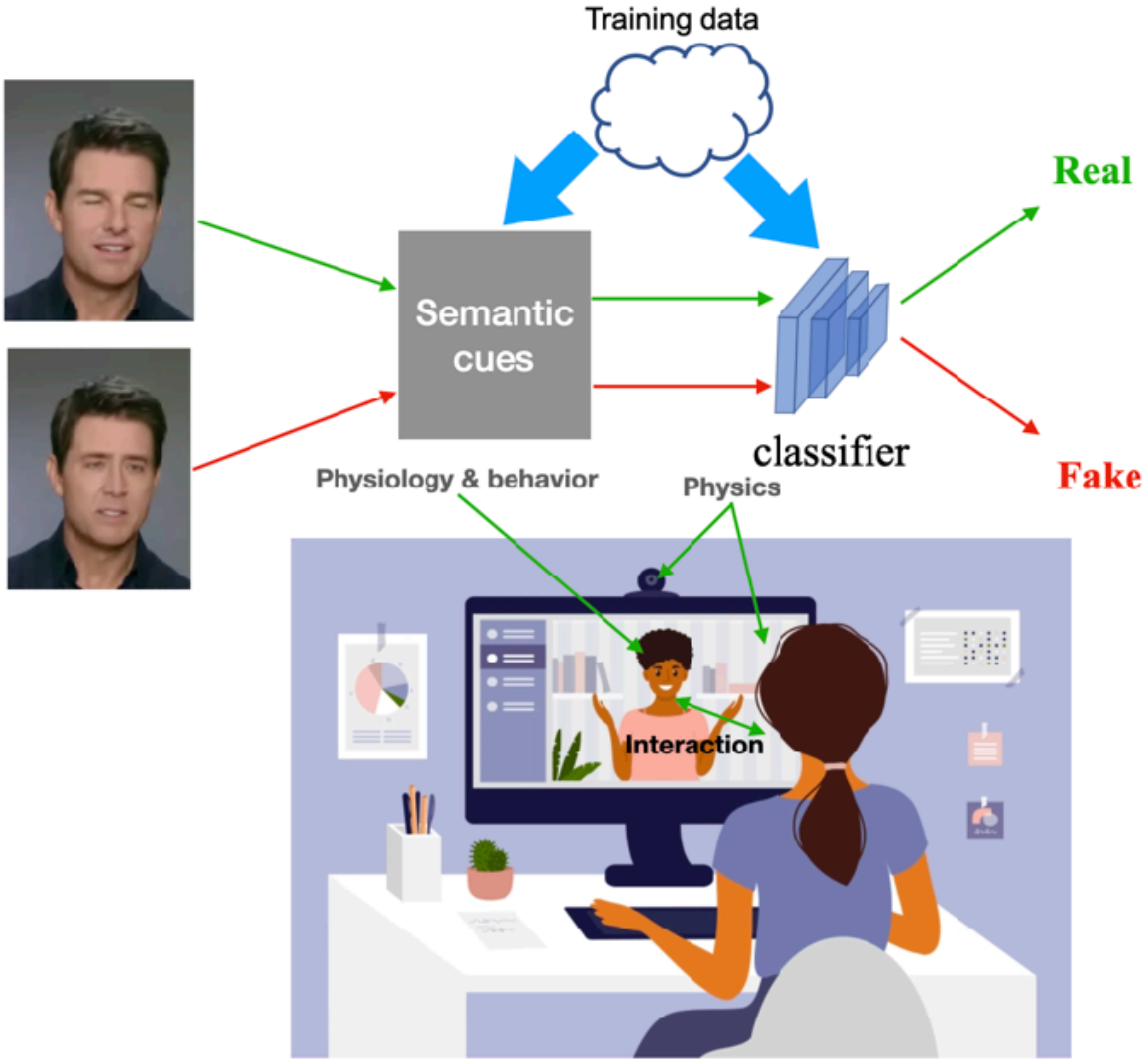
OM ▸ STATEMENTS AND RELE/

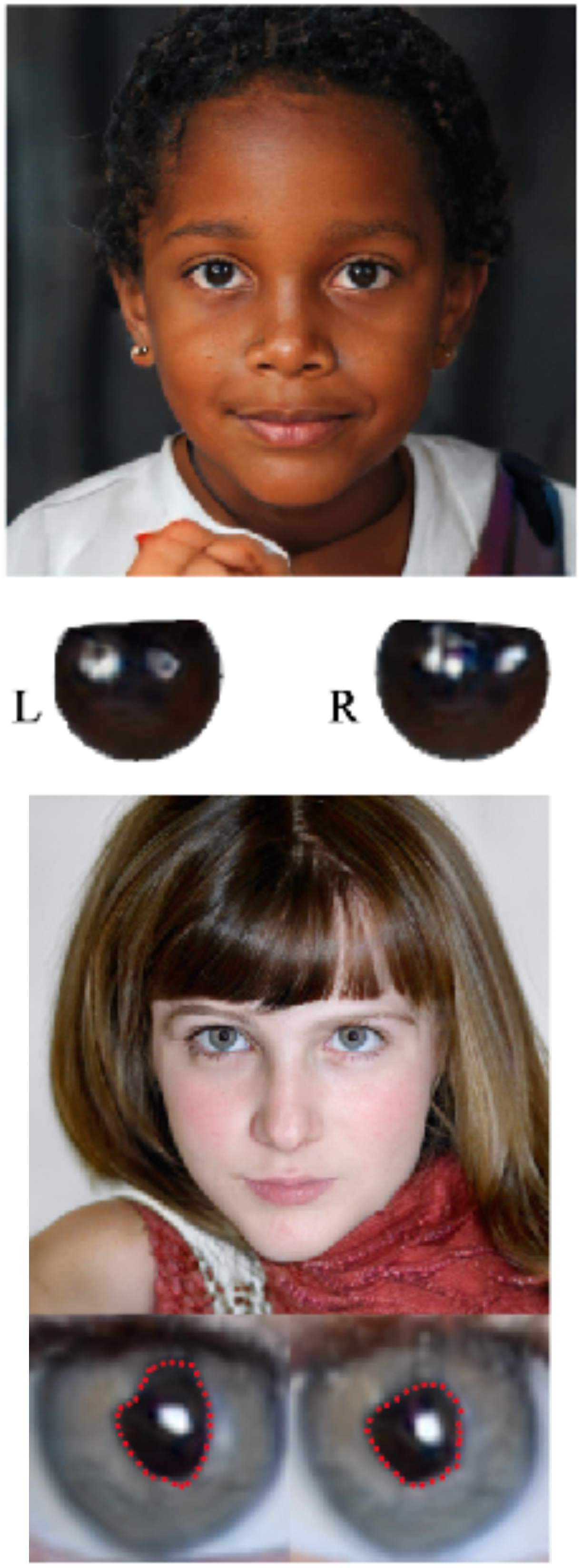
enabled fraud and
est practices for d
official content. Th

Watermarking

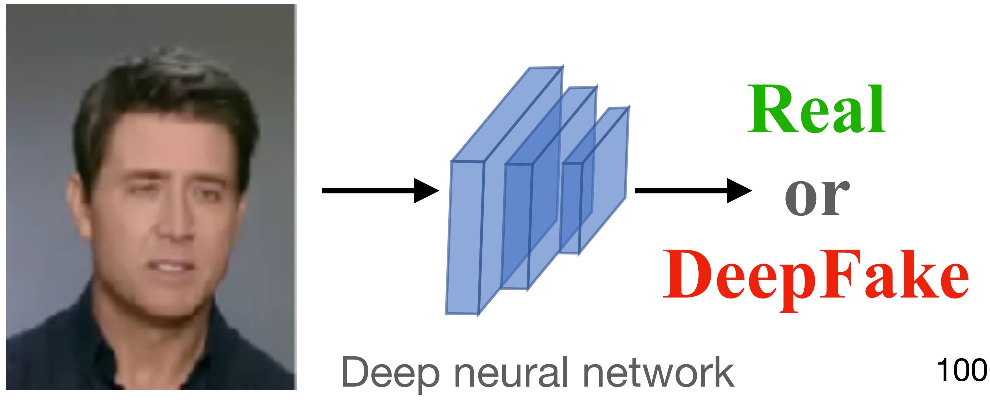
Commerce will develop guidance for content authentication and watermarking to clearly label AI-generated content. Federal agencies will use these tools to make it easy for Americans to know that the communications they receive from their government are authentic—and set an example for the private sector and governments around the world.

AIGC detection — semantic based





AIGC detection — data driven



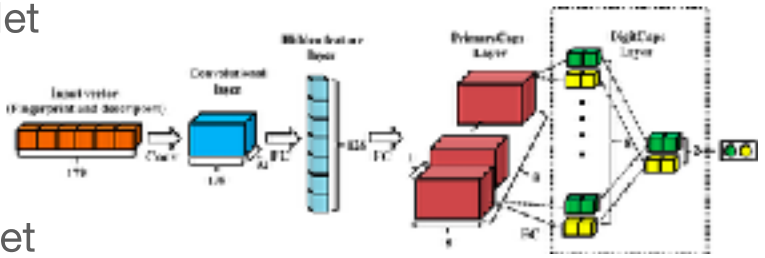
- More sophisticated and powerful NNs
- Larger and more diverse datasets
- Increasing performance on benchmarks

Resent

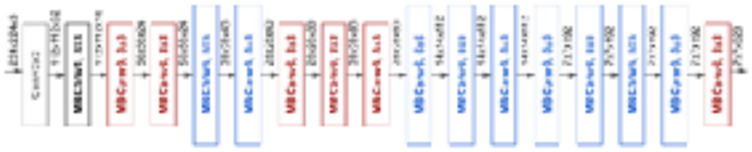
VGG

Xception Net

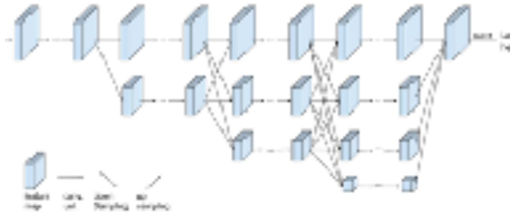
Capsule Net

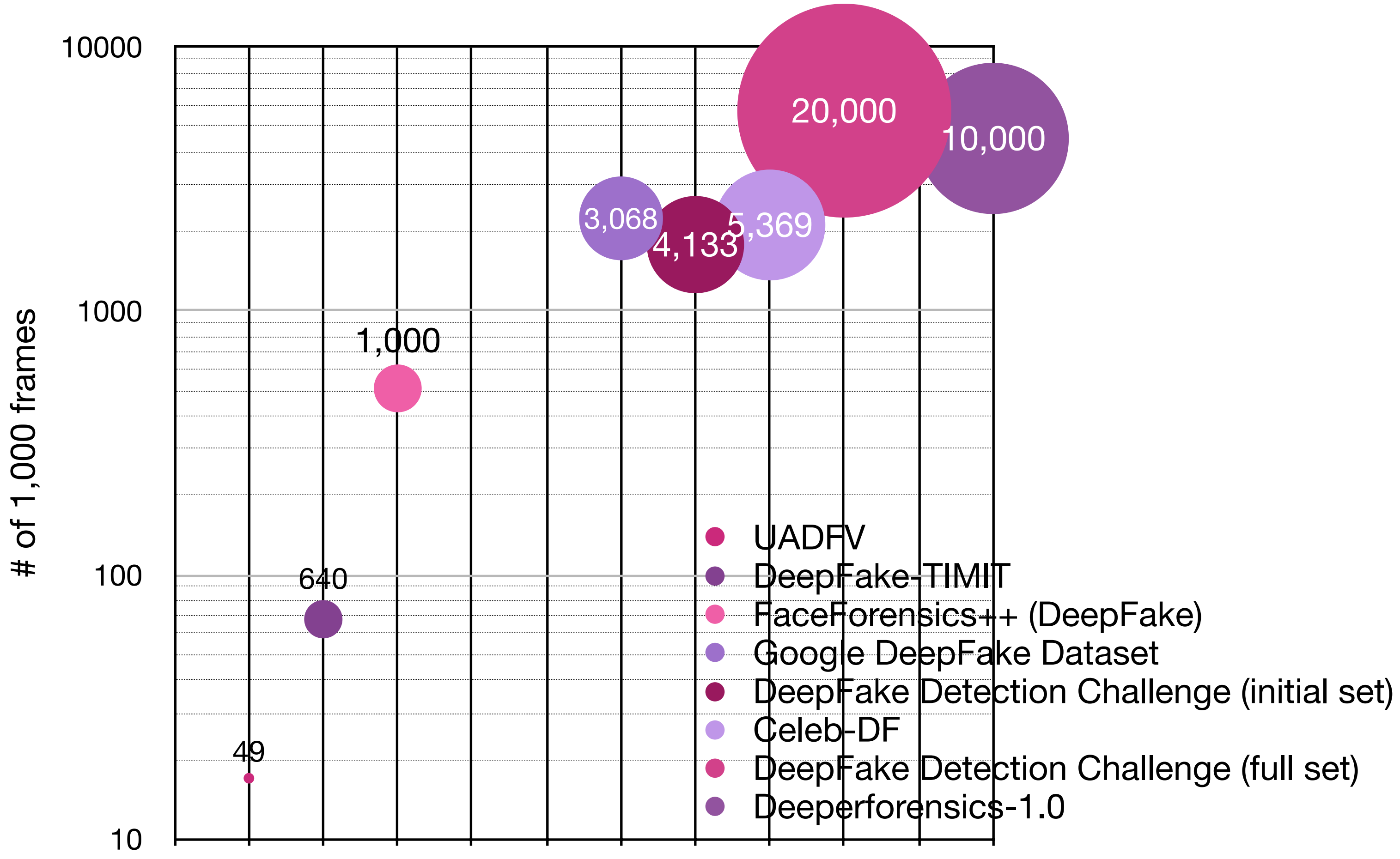


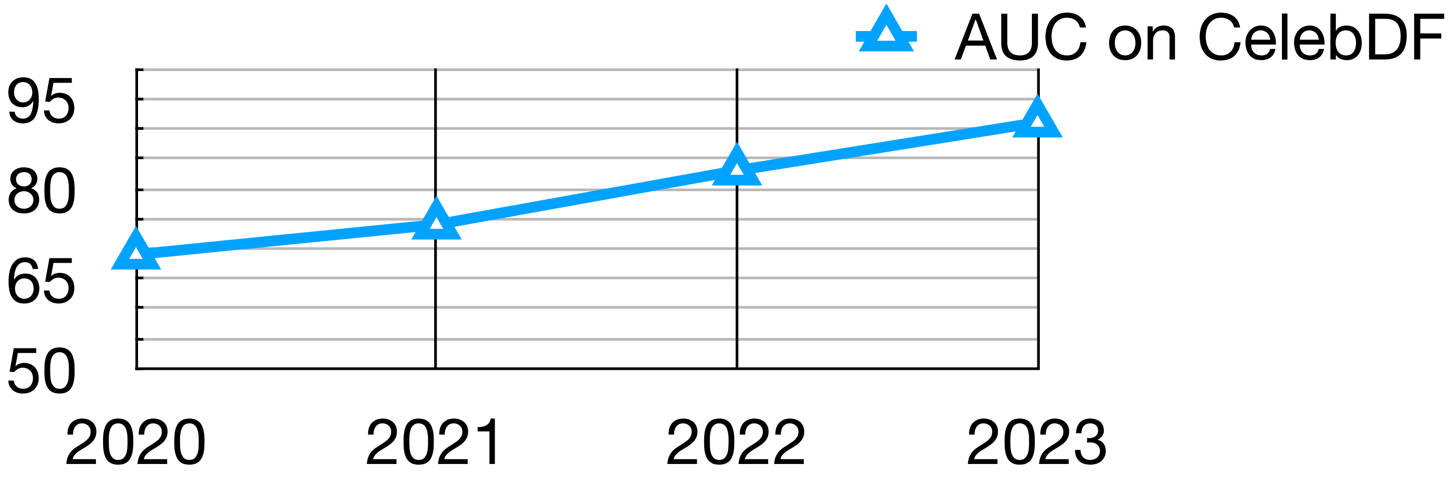
Efficient Net



HRNet




# of 1,000 frames
10000
1000
100
10
49
640
1,000
3,068
4,133
5,369
20,000
10,000
● UADFV
● DeepFake-TIMIT
● FaceForensics++ (DeepFake)
● Google DeepFake Dataset
● DeepFake Detection Challenge (initial set)
● Celeb-DF
● DeepFake Detection Challenge (full set)
● Deeperforensics-1.0



AIGC detection — are we there yet?

- Lack clarity in the problem definition
    - Is “real vs. DeepFake” well-defined?
- What is the purpose of detection
    - Use for triaging (overall accuracy/throughput) vs. use for evidence (individual accuracy/explanation)
- No option for “I am not sure”
    - Forcing detector to make decision on every input
- Too much reliance on labeled data and no diagnosis
    - Learning circumstantial attributes — source for lack of robustness to post-processing and low generalizability
- No user-friendly testbeds for performance
    - Commercial closed source systems vs. open source research code on code repositories

DeepFake-o-meter

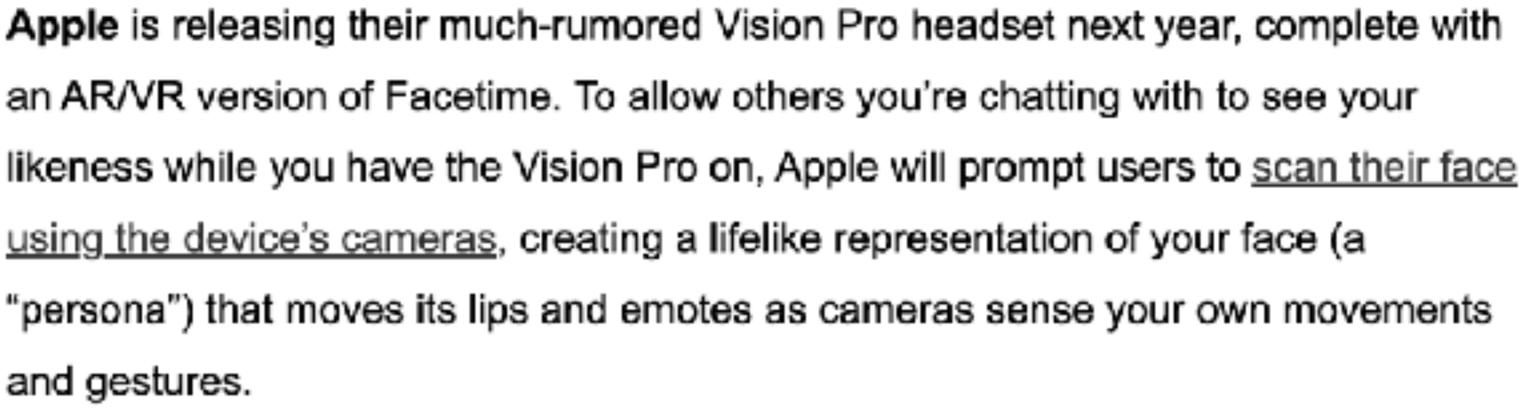
- Open-platform of open source AIGC detection methods
- Simple API/containerization to simplify integration
- Easy interface to casual users to run on individual media
- Comprehensive evaluation of in-the-wild performance

DEEPFAKE-O-METER
AN OPEN PLATFORM INTEGRATING STATE-OF-THE-ART DEEPFAKE DETECTION METHODS

https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/

AIGC detection — what’s next?

- Detection of AIGC may become irrelevant

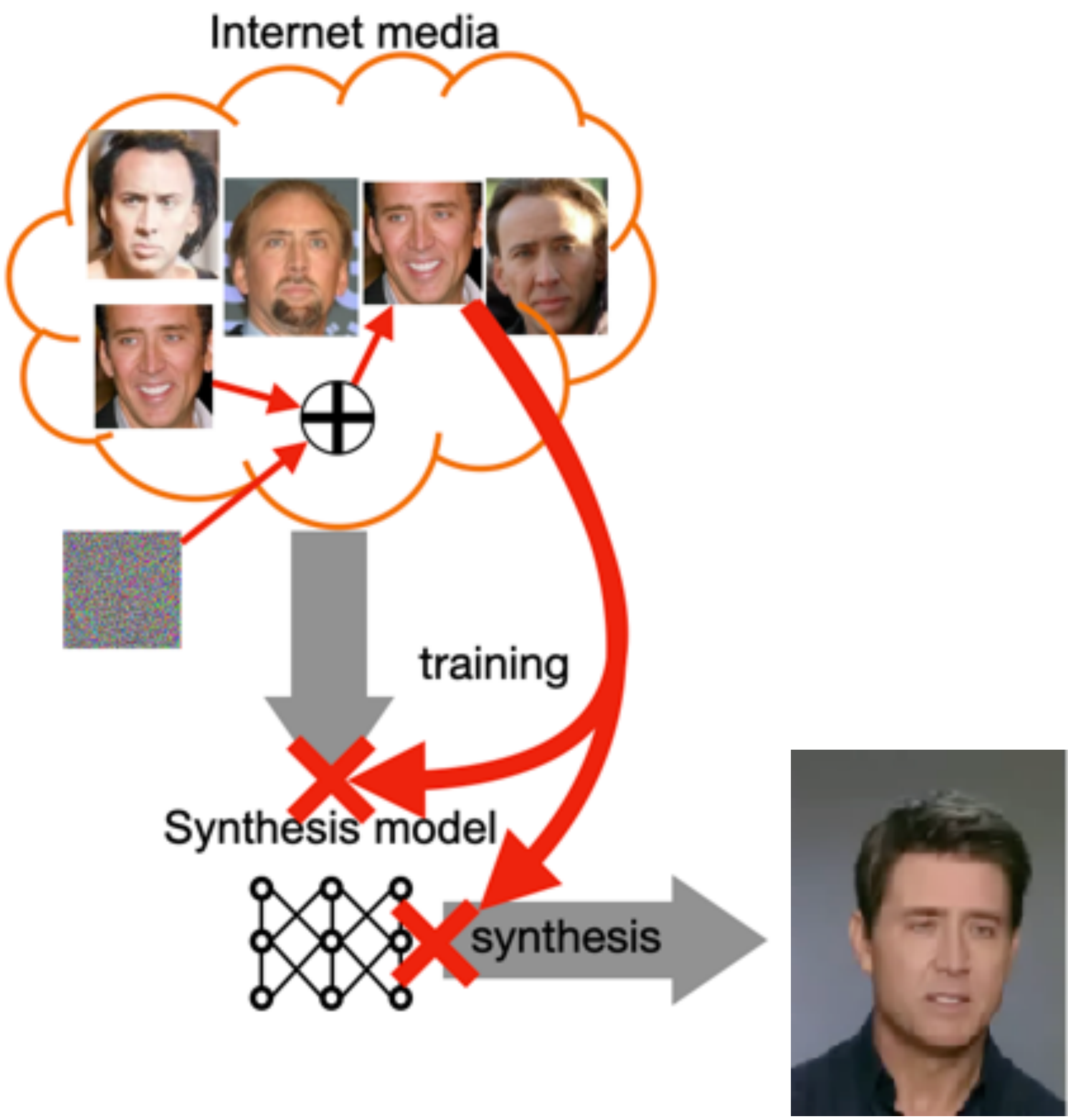


Apple is releasing their much-rumored Vision Pro headset next year, complete with an AR/VR version of Facetime. To allow others you're chatting with to see your likeness while you have the Vision Pro on, Apple will prompt users to scan their face using the device's cameras, creating a lifelike representation of your face (a "persona") that moves its lips and emotes as cameras sense your own movements and gestures.

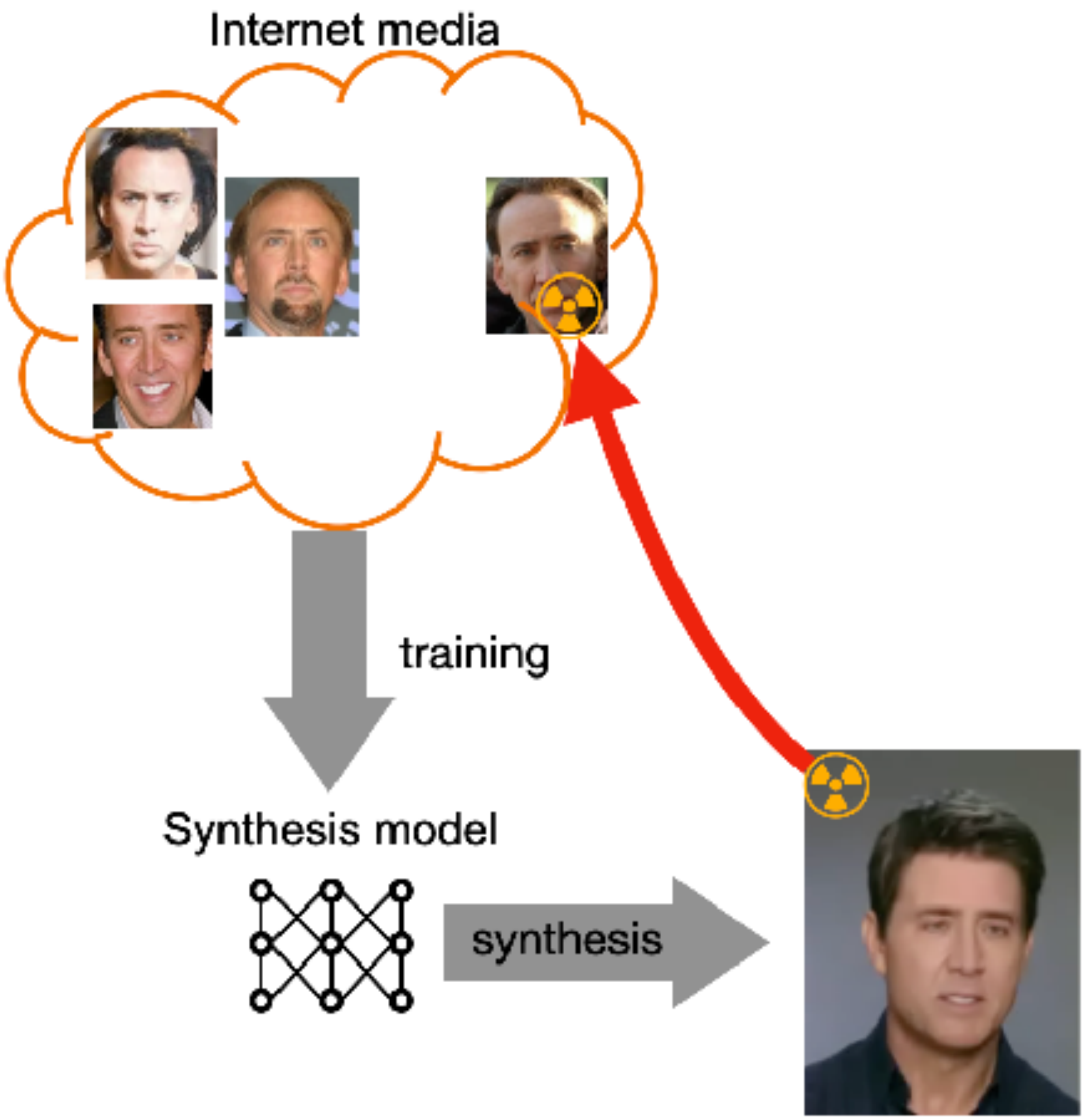
  - Verification of authenticated use of gen AI models
    - Liveness/presence detection for real-time AIGC
- Attribution and provenance of the generation process
- Providing fine-grained analysis — spatial and temporal localization
- Other modalities and cross-modality analysis
- Handle the zero-day attack: making detection generalizable

Beyond detection

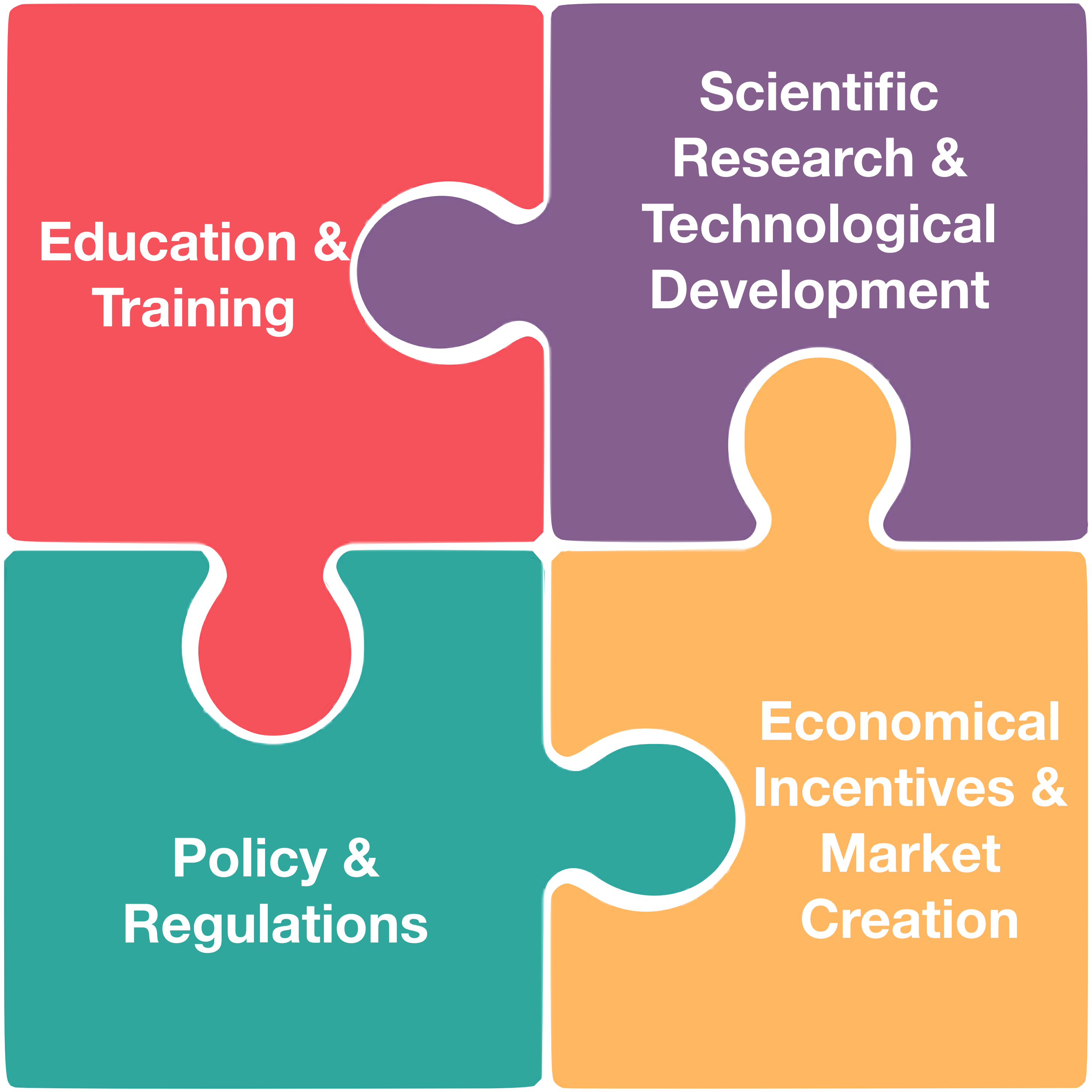
Obstructing DeepFakes



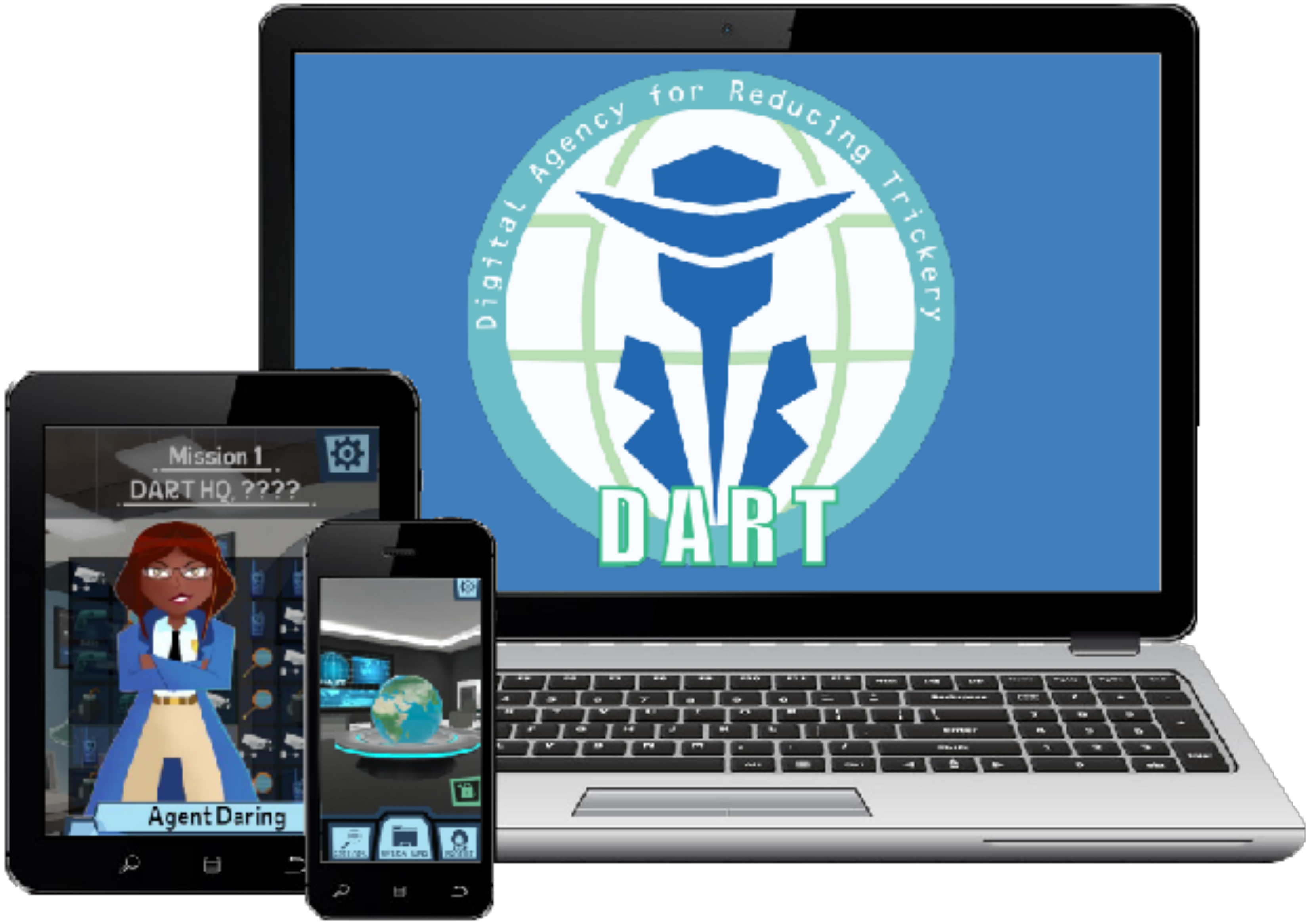
Tracing DeepFakes



Beyond technologies


Education & Training
Scientific Research & Technological Development
Policy & Regulations
Economical Incentives & Market Creation

DART Learn



DeepCover

Conclusions

- AIGC continues to have rapid developments
- Misused AIGC (aka DeepFakes) erode our trust to online information
- There is no “silver bullet” solution, we need all
  - Content authentication and provenance
  - Watermarking
  - Reactive forensics — detection, attribution
  - Active forensics — affecting training/synthesis
- Future directions for AIGC detection
  - Improve robustness, transparency, and explainability
  - Handling diversity and complexity of generation
  - Providing more insights beyond binary classification

Thank you!

Contact: siweilyu@buffalo.edu