Visual Content Verification
WP3 – T3.2 What we have done in 2022

Sohail Ahmed Khan, Duc Tien Dang Nguyen

Research Centre for Responsible
Media Technology and Innovation
Project number 309339

How can we assist media industry in verifying deepfakes and cheapfakes efficiently and effectively?

RQ1: How multimedia content verification is carried out in the media industry?

SRQ 1.1
What is the State of the Art in multimedia verification in journalism?

SRQ 1.2
How media practitioners verify content using automated tools and the limitations of the available tools?

SRQ 1.3
What are the limitations of the current tools and software available for multimedia content verification?

RQ2: How can we fight deepfakes within the news domain?

SRQ 2.1
What is the State of the Art in deepfake detection?

SRQ 2.2
How can we make the deepfake detection systems more robust and generalizable?

SRQ 2.3
How can we address and improve deepfake detection in the NEWS domain?

RQ3: How can we fight cheapfakes within the news domain?

SRQ 3.1
What is the State of the Art in cheapfake detection?

SRQ 3.2
Separate models for image and text data, or a single model for both? What is better?

SRQ 3.3
How can we address and improve cheapfake detection in the NEWS domain?

RQ1. **How visual UGC verification is carried out**

Media Futures

Visual User-Generated Content Verification in Journalism

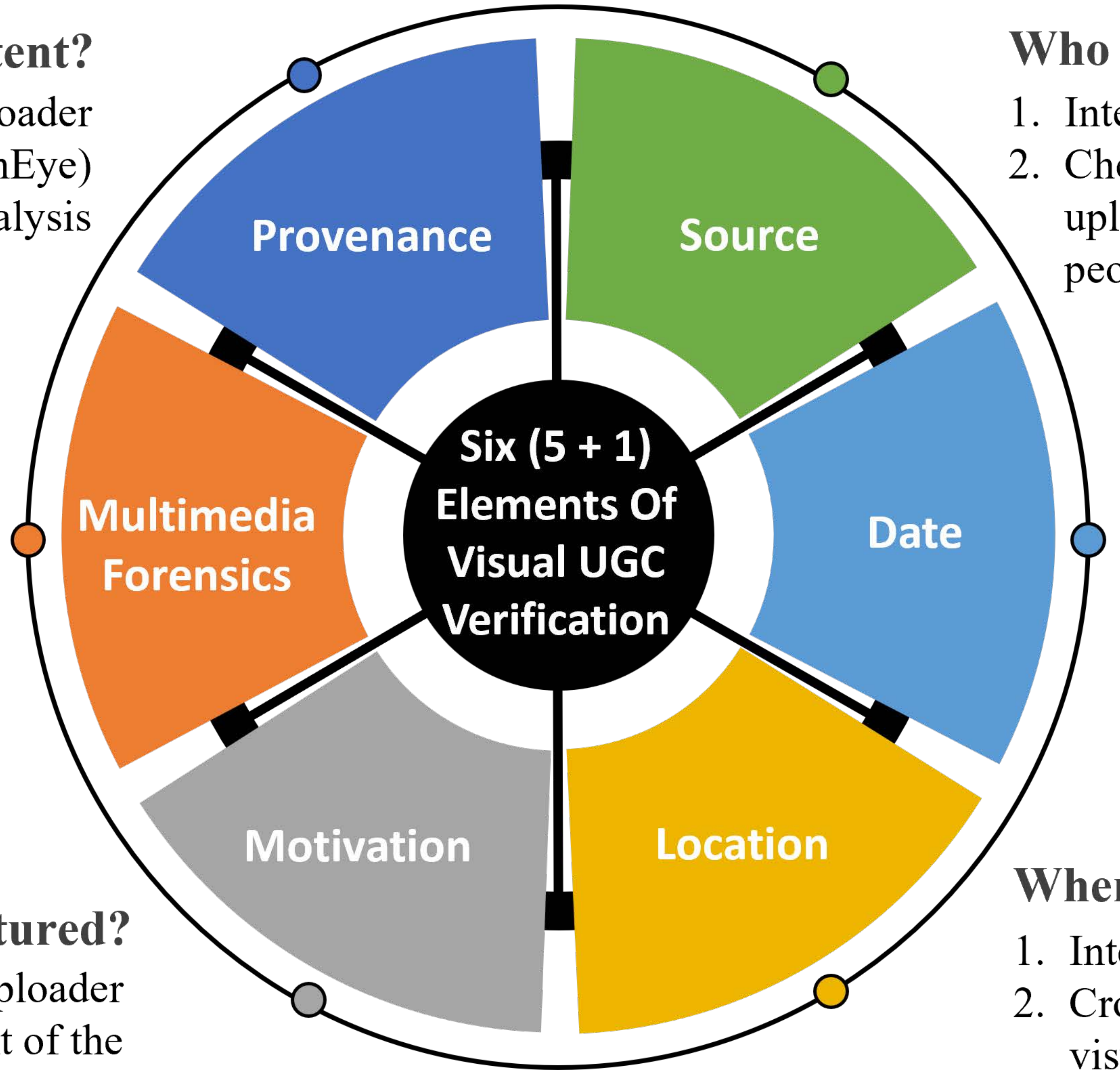
Sohail Ahmed Khan, Ghazaal Sheikhi, Andreas Lothe Opdahl, Fazle Rabbi, Sergej Stoppel, Christoph Trattner, Duc-Tien Dang-Nguyen

Is this the original piece of content?

1. Interviewing the uploader
2. Reverse image search (Google, TinEye)
3. Exif metadata analysis

Who originally captured the content?

1. Interviewing the uploader
2. Checking the digital footprint of the uploader e.g., social media analysis, people search engines



Is the content manipulated or synthetic?

1. Analysing visuals for manipulations using available image/video forensics tools
2. Analysing visual content to check if it's manipulated or generated using deep networks

When was the content originally captured?

1. Interviewing the uploader/eyewitness
2. Exif metadata analysis
3. Reverse image search

Why was the content captured?

1. Interviewing the uploader
2. By looking at the digital footprint of the source person i.e., analysing his/her social media accounts, or other online presences

Where was the content captured?

1. Interviewing the uploader/eyewitness
2. Cross-checking location shown in the visual content using online maps
3. Social media analysis

Visual User-Generated Content Verification in Journalism

Sohail Ahmed Khan, Ghazaal Sheikhi, Andreas Lothe Opdahl, Fazle Rabbi, Sergej Stoppel, Christoph Trattner, Duc-Tien Dang-Nguyen

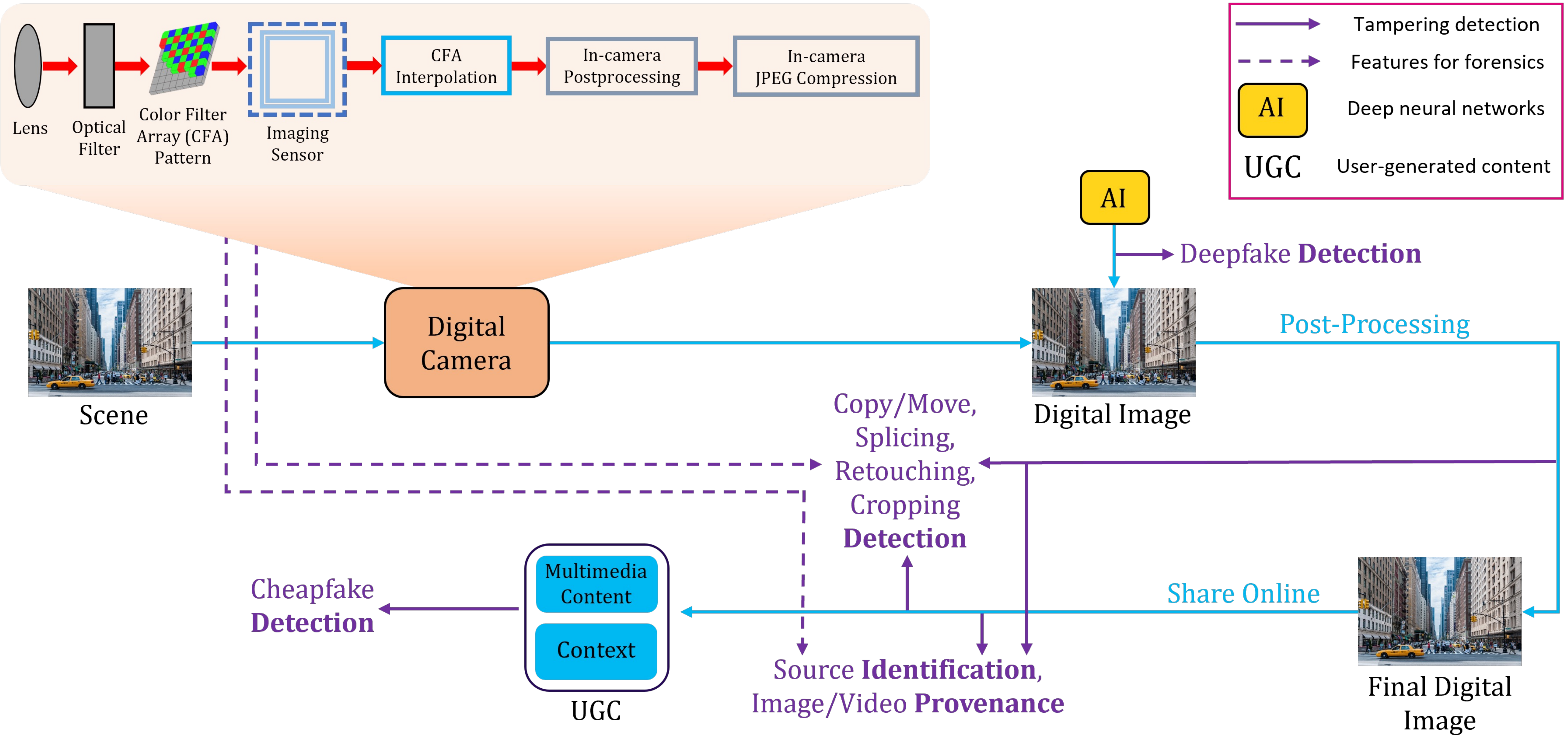


TABLE 1. A summary of multimedia problems presented in Section III. We also list suitable forensics techniques, as well as available tools to detect/debunk these forgeries. Some of content in this table is inspired from [8]. An analysis of the tools can be found in Table 3.

Modification Category	Problem	Examples from the News Domain	Forensics Techniques	Tools
Similar	Source Identification	A video game clip was mis-captioned and shared on social media platforms in the context of Russian invasion of Ukraine. The computer generated clip claimed to show "Ghost of Kyiv", a fictitious Ukrainian fighter pilot shooting down a Russian fighter jet [31].	Source identification is carried out by analysing metadata information, CFA interpolation patterns, sensor noise fingerprints, JPEG compression artifacts. Deep CNN models have also been employed for the source identification task.	MeVer Image, InVID, Ghiro, FotoForensics, Forensically, DeDigi, Online Exif Viewer, exifdata, YouTube Data-Viewer
	Image/Video Provenance	An image went viral on social media in 2021 claiming to show a heart-shaped sunset over a beach. The image was found to be mis-captioned, and the original image (digital artwork) was actually posted on Instagram by a user in 2020 [32].	For provenance analysis metadata information, noise fingerprints, DCT features are used to train statistical models. Deep learning models are also proposed for provenance analysis.	MeVer Image, InVID, Ghiro, FotoForensics, Forensically, DeDigi, Google/TinEye Image Search
Enhanced/ Retouched	Retouching	US President Donald Trump's official Facebook and Instagram handles shared his edited photos to show him with a tightened waistline, elongated fingers, a slimmed neck and shoulder, higher crotch and tightened hair [8].	Retouching forgeries are typically detected using noise patterns, histogram analysis. Deep CNN models are also used to detect these forgeries.	MeVer Image, InVID, Ghiro, FotoForensics, Forensically, DeDigi, Google/TinEye Image Search
	Cropping	During the inauguration ceremony of US President Donald Trump, the White House cropped official photos in a way that made the crowd seem larger. For reference, see Figure 3.	Cropped images are normally multiple compressed, they can be detected by analysing the image compression qualities, image histogram, or blocking artifacts. Deep learning models are also proposed to detect image cropping.	MeVer Image, InVID, Ghiro, FotoForensics, Forensically, DeDigi, Google/TinEye Image Search
Doctored	Copy-Move	Sepah News, owned by Iran's Revolutionary Guards posted forged images using copy/move forgery to show four missiles, instead of the original 3. The image was edited by copying and pasting one of the missiles from the original image itself [20].	Two widely used detection methods are, (1) Block matching based method exploiting DCT and DWT features; and (2) Key-point matching based methods exploiting SIFT, SURF features to detect manipulated images. Some approaches use deep learning models as well.	MeVer Image, InVID, Forensically, Google/TinEye Image Search

TABLE 3. A list of useful tools for visual UGC verification, and some of their limitations. The associated visual UGC verification elements described in Section II are also presented in this table, where 1 = **Provenance**, 2 = **Source**, 3 = **Date**, 4 = **Location**, 5 = **Motivation** and 6 = **Multimedia Forensics**.

Tool	Use Case	Element	Limitations
WeVerify - InVID https://tinyurl.com/mtfcj59s	Image/Video Analysis, Metadata Analysis, Frame Extraction	1, 3, 4, 6	Struggles against heavy compression, requires some level of training to be used.
TrulyMedia https://www.truly.media/	Contextual Image/Video Analysis, Identity Verification	1, 2, 3, 4, 5	Restricted access, no forensics tools are made available, no documentation available.
MeVer https://caa.iti.gr/	Contextual Visual Content Analysis, Metadata Analysis	1, 2, 3, 4, 5, 6	Relies heavily on the already available information on the web, not useful when there is no related information available about fairly recently surfaced fake visual content.
FotoForensics http://fotoforensics.com/	Image Analysis, Metadata Analysis, String Extraction	1, 3, 4, 6	No dedicated copy-move detector, struggles against heavy compression, does not allow customized forensics filters.
Forensically https://29a.ch/photo-forensics/	Image Analysis, Metadata Analysis, String Extraction	1, 3, 4, 6	Struggles against heavy compression, requires some level of training to be used.
Ghiro https://www.imageforensic.org/	Image Analysis, Metadata Analysis, GPS Localization	1, 3, 4, 6	No copy-move detector, struggles against heavy compression, does not allow customized forensics filters.
DeDigi http://www.dedigi.tech/	Image Analysis, Metadata Analysis, GPS Localization	1, 3, 4, 6	Struggles against heavy compression, user-interface can be improved.
Deepware https://deepware.ai/	Deepfake Detection	6	Only analyzes videos with duration of less than 10 minutes, the available deepfake detection models can be improved.
Snopes https://www.snopes.com/	Fact Checking	1, 2, 3, 4	Only helps if the image/video being verified has already been fact-checked.
Google Image Search https://www.google.com/imghp	Reverse Image Search	1, 2, 3	Will not help if the visual UGC being verified has been shared for the first time, or fairly recently.

Media Forensics.

Home Image Forensics Publications Tools Contribute About

Keywords Search

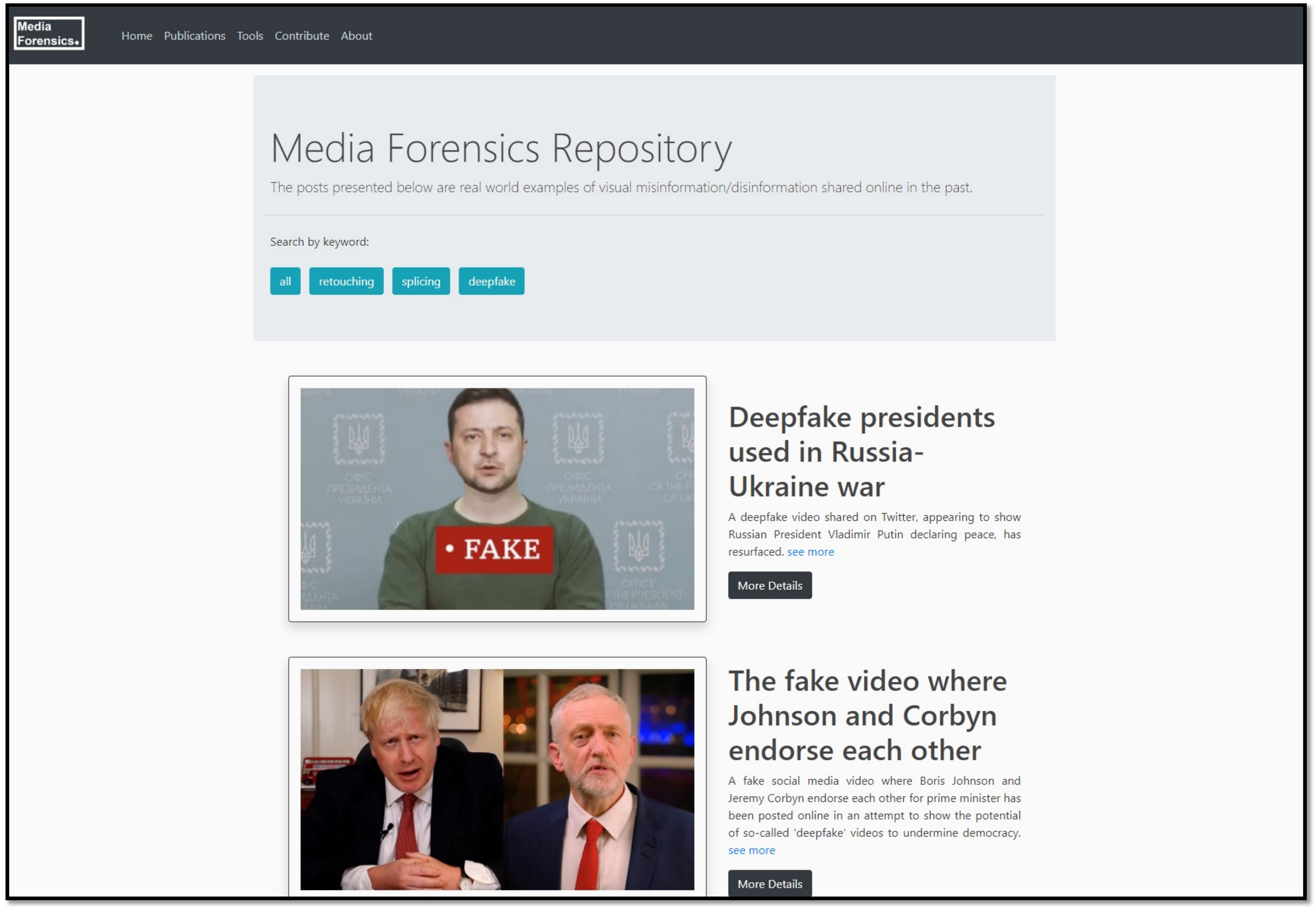
Click on a keyword to search relevant tools based on a specific category.

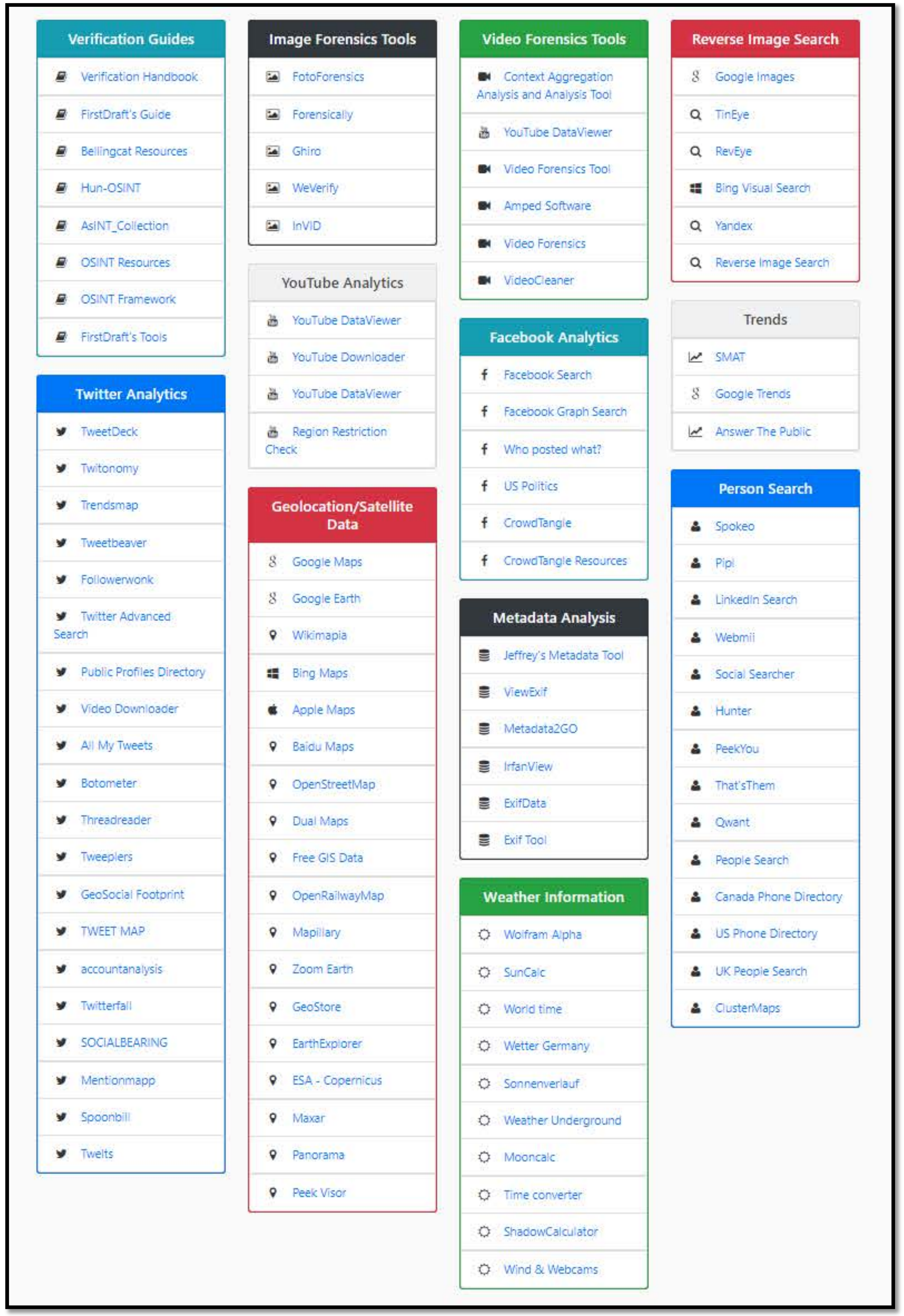
all | reverse-image-search | shadow-analysis | metadata-analysis | image-forensics | facial-recognition | person-identification | face-restoration | image-classification | image-organization | image-editor | ipvm-calculator | deepfake-detector | misc

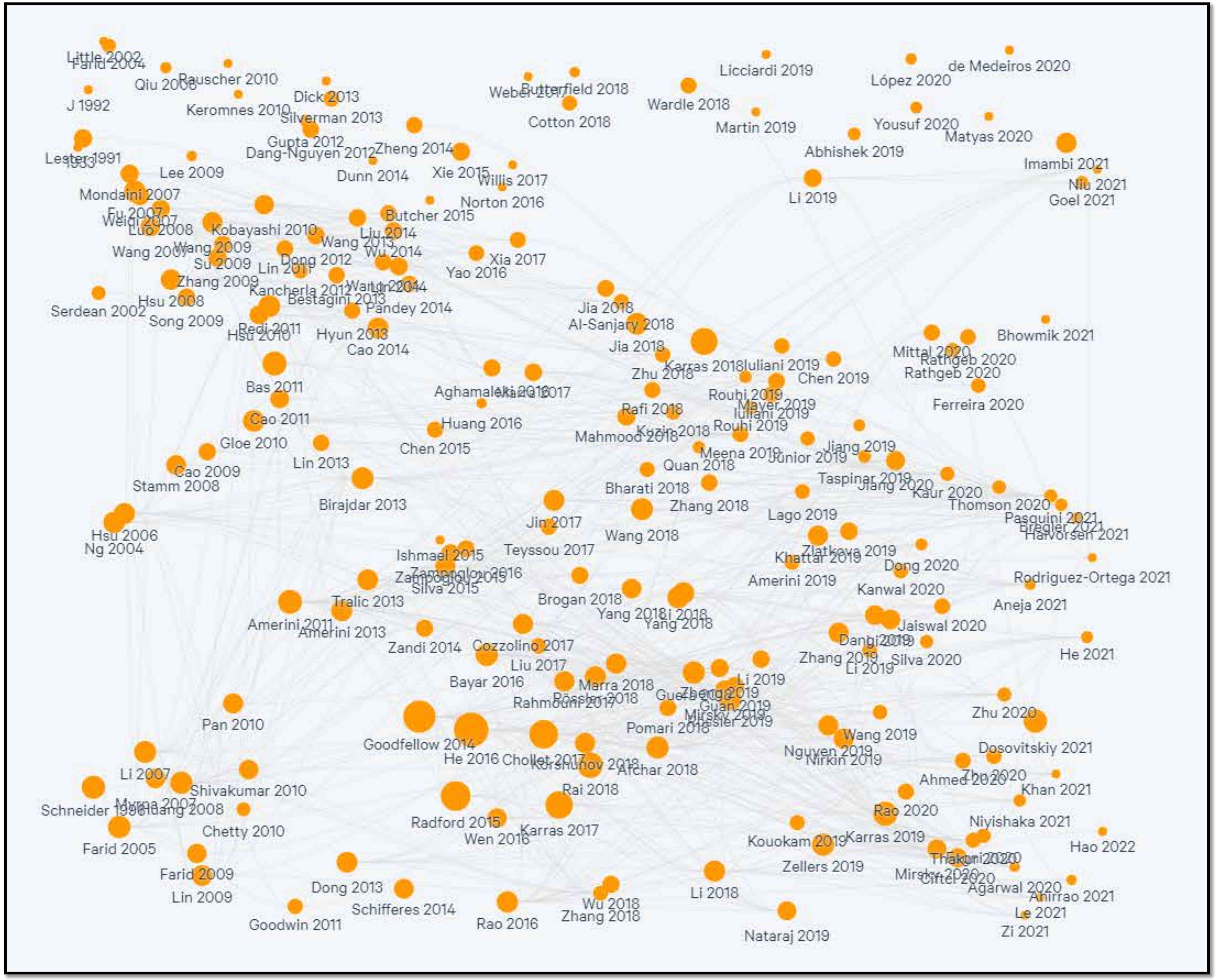
#	Type	Name	Category	Description	Subscription	Guide
1	Reverse image search	**InVID-WeVerify**	Image/Video Verification	Verification plugin to help journalists verify images and videos. Contextual data, Metadata, reverse search (Google, Yandex, Baidu), image forensic, Magnifier). @WeVerify on Twitter.	Free	Guide To Using Reverse Image Search For Investigations
2	Reverse image search	**Google Lens**	Image/Video Verification	Google Lens but in your browser - it's better than Google Image reverse search. h/t @Henkvaness	Free	
3	Shadow analysis	**SunCalc**	Image/Video Verification	Make an approximation of the time of the day using shadow direction.	None	Using the Sun and Shadows for Geolocation
4	Metadata analysis	**metadata2go**	Image/Video Forensics	Check metadata for both photos and videos online.	Free	
5	Metadata analysis	**Reveal Image Verification Assistant**	Image/Video Forensics	Forensic providing eight filters to detect still images alterations. Web-based image tool. Also available within InVID verification plugin.	Free	
6	Metadata	**ExifPurge**	Image/Video	EXIF Purge is a small portable application to remove	Free	

Multimedia Forensics Repository

- ❑ A web repository containing diverse set of resources, e.g.,
  - Tools and resources helpful for journalists/fact-checkers to verify visual user-generated content found online
  - Visual content forgeries widely spread online in the past
  - Research publications focusing on detecting multiple different forms of visual content forgeries
- ❑ We aim at updating this repository with new resources over the next years
- ❑ We plan to incorporate automated visual content verification demos (for cheapfake and deepfake media detection) within this repository in the future



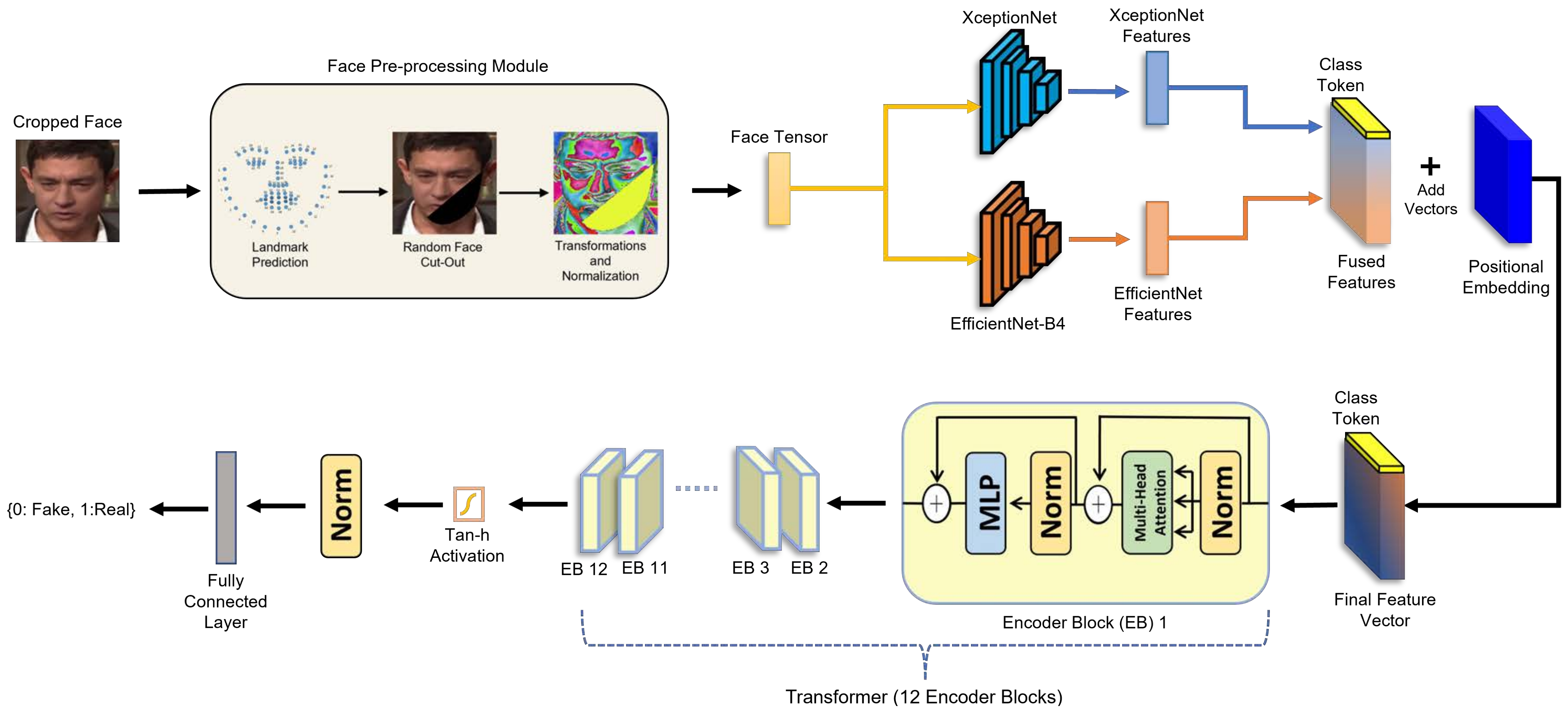




Media Futures

RQ2. **Deepfake Detection**

Hybrid Transformer Network for Deepfake Detection



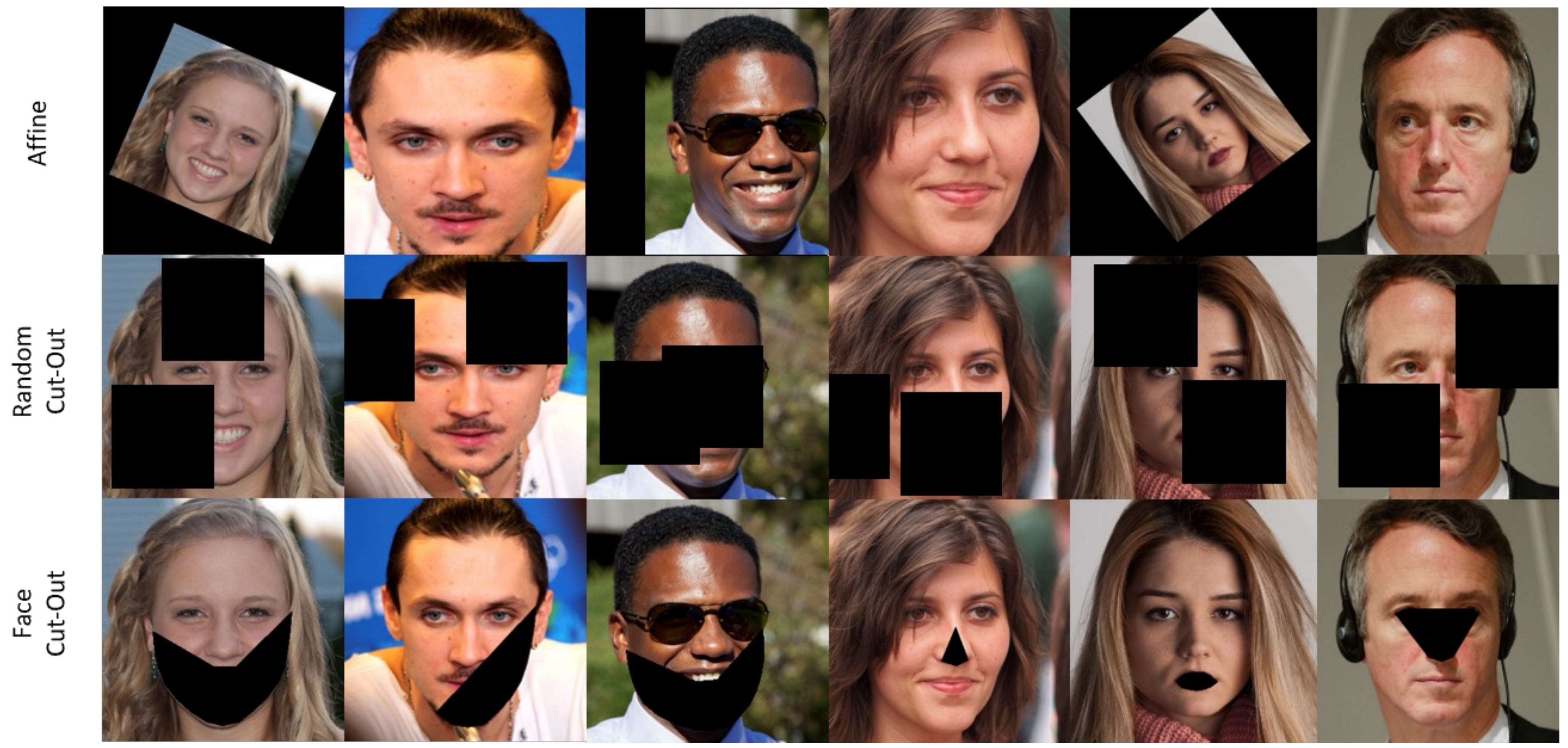


Table 1: Performance (accuracy) comparison of a number of different deepfake detection baseline models on FaceForensics++ dataset. Each of the mentioned model was trained on all subsets of the FaceForensics++ dataset at once. Best results are highlighted.

Approach	Deepfakes	Face2Face	FaceSwap	NeuralTextures	Pristine	Cumulative
Steg. Features + SVM [11]	68.80%	67.69%	70.12%	69.21%	72.98%	70.97%
Cozzolino *et al.* [6]	75.51%	86.34%	76.81%	75.34%	78.41%	78.45%
Bayar and Stamm [2]	90.25%	93.96%	87.74%	83.69%	77.02%	82.97%
Afchar *et al.* [1]	89.55%	88.60%	81.24%	76.62%	82.19%	83.10%
Rossler *et al.* [22]	97.49%	97.69%	96.79%	**92.19%**	95.41%	95.73%
Qi *et al.* [21]	**99.70%**	**98.90%**	97.80%	-	-	-
Ours (Face cut-out)	97.85%	97.85%	96.42%	90.71%	95.00%	95.57%
Ours (Random cut-out)	98.57%	98.57%	**97.85%**	92.14%	**97.85%**	**97.00%**

A simple straightforward architecture that uses much less training data while keeping comparable results to other more advanced state-of-the-art approaches

Khan, S.A. and Dang-Nguyen, D.T., 2022, September. Hybrid Transformer Network for Deepfake Detection.
In *International Conference on Content-based Multimedia Indexing* (pp. 8-14) **Paper Award Runner-up**

On-going work: Deepfake Detection: A Comparative Analysis

1. Introduction
2. Literature Review
3. Methodology
3.1. Datasets
3.1.1. FaceForensics++
3.1.2. FakeAVCeleb
3.1.3. DFDC
3.2. Preprocessing and Augmentations
3.3. Models
3.3.1. Image Models

- Xception Net
- Res2Net-101
- EfficientNet-B7
- Vision Transformer (ViT Base)
- Swin Transformer (Swin Base)
- Multiscale Vision Transformer (MViT V2 Base)

3.3.2. Video Models

- ResNet 3D
- Video Swin Transformer
- Multiscale Video Transformer

3.4. Training Strategies
3.4.1. Supervised
3.4.2. Self-Supervised

- BYoL
- MoBY
- Dino

"DON'T BELIEVE EVERYTHING YOU AND A DEEPFAKE DETECTOR SEE"

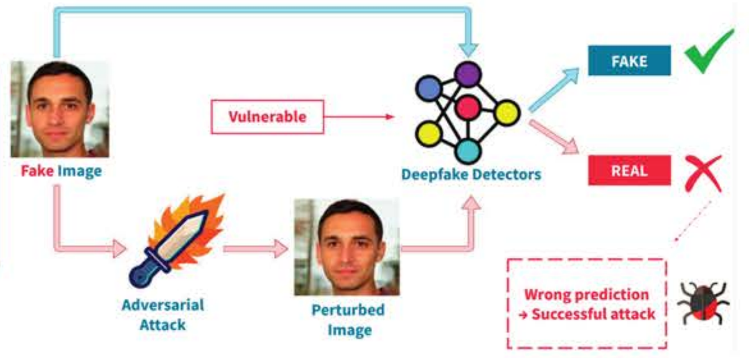
ADVERSARIAL ATTACKS ON DEEPFAKE DETECTORS: A Practical Analysis

Ngan Hoang Vo [1,2], Khoa D. Phan [1,2], Anh-Duy Tran [1,2] & Duc-Tien Dang-Nguyen [3,4]
[1] FIT, University of Science, HCMC, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
{vhngan,pdkhoa}@apcs.fitus.edu.vn, taduy@fit.hcmus.edu.vn
[3] University of Bergen, Norway
[4] Kristiania University College, Norway
ductien.dangnguyen@{uib.no, kristiania.no}

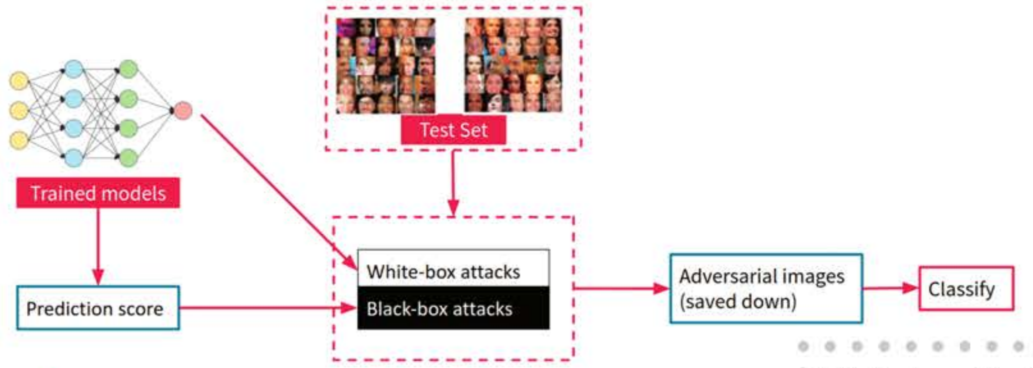
OVERVIEW

Deepfake creates fake images which are indistinguishable from real ones by the human eye. In practice, most detection methods use simple deep neural networks (DNNs) as the backbone. However, they are vulnerable to adversarial examples. This work presents practical pipelines in both white-box and black-box attack scenarios that can fool DNN-based Deepfake detectors into classifying fake images as real.

KEY IDEA



EXPERIMENTS



Different backbone, same Deepfake detection pipeline.

Goal: Successfully modifying a fake/non-fake image so that the detection models will give wrong predictions.

Lower accuracy is better!

Backbone Models:

- Baseline
- XceptionNet
- EfficientNet B3
- EfficientNet v2s

White-box attacks: unrealistic.
Black-box attacks: need to invoke the detector many times.
→ Transfer Attacks:
Use white-box attacks in black-box settings.

Adversarial Attacks:

White-box:

- Fast gradient sign (FGS) [Goodfellow et al., 2014]
- Iterative gradient Sign (IGS) [[Kurakin et al.,2017]
- Iterative Gradient Sign with data augmentation (IGS \w aug)
- Deepfool [Moosavi-Dezfooli et el.,2016]

Black-box:

- Simple Black-box Attack (SimBA) [[Guo et al.,2019]
- Natural Evolution Strategy (NES) [[Wierstra et al.,2008]

RESULTS

Accuracy <1%
IGS is our best white-box attack

Among black-box attacks
NES is better than SimBA

Client-side
White-box attacks can be performed when the detection model is deployed in client-side.

Too many queries
are required to perform black-box attacks.

Real threat
Adversarial attacks pose a very practical threat to Deepfake detectors as well as real life.

Defensive approaches
The adversarial attack is a clever way to do pressure testing and debugging on machine learning models that are considered mature before they are actually being deployed in the field.

Download the paper & more information

RQ3. **Cheapfake Detection**

ACM Multimedia Grand Challenge on Detecting Cheapfakes

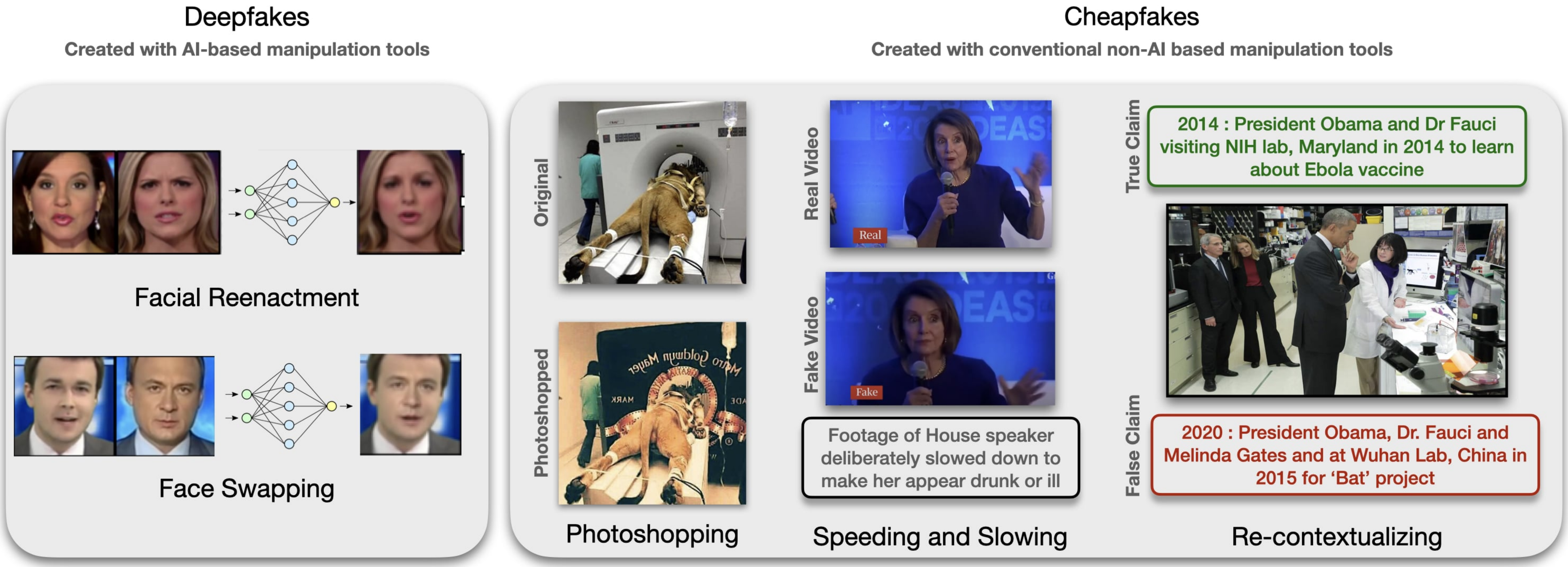


Figure 1: Deepfakes (left) are defined as falsified media created using sophisticated AI-based media manipulation tools and techniques. Cheapfakes (right) include falsified media created with/without contemporary non-AI based editing tools which are easily accessible. Photoshopping tools can be used to tamper with images. Videos can be sped up or slowed down to change the intent or misrepresent the person in the video. Re-contextualizing includes associating falsified or unrelated claims with a genuine image to misrepresent events or persons. This challenge is focused on detecting re-contextualized cheapfakes. Image sources: [6, 12, 13, 22, 27]

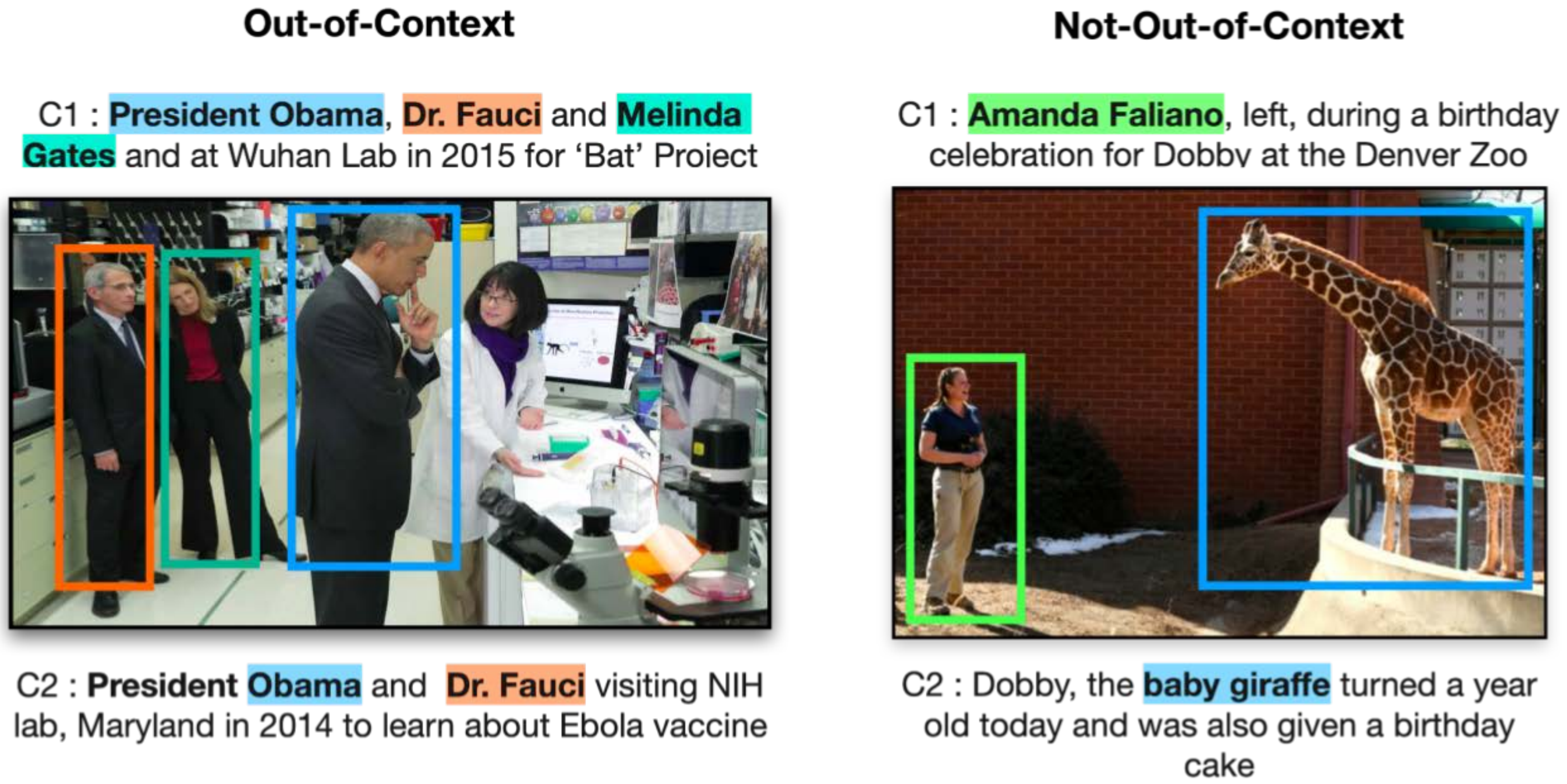


Figure 2: Each image in the dataset is accompanied by one or two captions that the image was circulated together with on the Internet. On the left, one of the two captions is misleading with an alteration of context, indicating out-of-context (OOC) misuse. On the right, none of the two captions are misleading, hence not-out-of-context (NOOC). Image source: [3]

Related Master Projects

Related Master Projects

- Vegard Velle Sjøen (graduated June 2022): Digital Image Forensics In The Wild: Social Media Platforms
- Espen Bøe (being submitted, December 2022): Measuring Engagement from Interactive Design Methods in Non-Fiction News Articles
- Eivind Moholdt (on-going, June 2023): Out of context <image, text> pairs detection
- Erik Gjertsen (on-going, June 2023): Detecting images generated by Dall-E

Media Futures

Collaborations

Related Projects

- EU - NORDIS (2021 - 2023): Nordic Observatory for Digital Media and Information Disorder
- NFR - NewsAngler (2018 - 2022): Finding new and unexpected angles on unfolding news stories, along with suitable background information
- IPN - EXPLAIN (2022 - 2024): Explainable AI for automated fact-checking and additional insight - when facts are not enough

Use of fact-checking tools

Task	Tool	Users (%, N = 17)
Audio transcription	Amberscript	11,8
	oTranscribe	5,9
Image and video verification	TinEye	41,2
	InVID	29,4
	Google Image	23,5
	Citizen Evidence	5,9
	PimEyes	23,5
	Deepware	5,9
Geolocation	Google Earth	17,6
	Google Street View	5,9
	Google Maps	11,8

Task	Tool	Users (%, N = 17)
Searching & verifying	OSINT Tools	35,3
	Google	52,9
	Google Cache	5,9
	WayBack Machine	41,2
Social networks monitoring	CrowdTangle	70,6
	Storyboard.news	23,5
	Twitter Advanced Research	5,9
	TweetDeck	23,5
Translation	Google Translate	11,8

Reported on May 2022

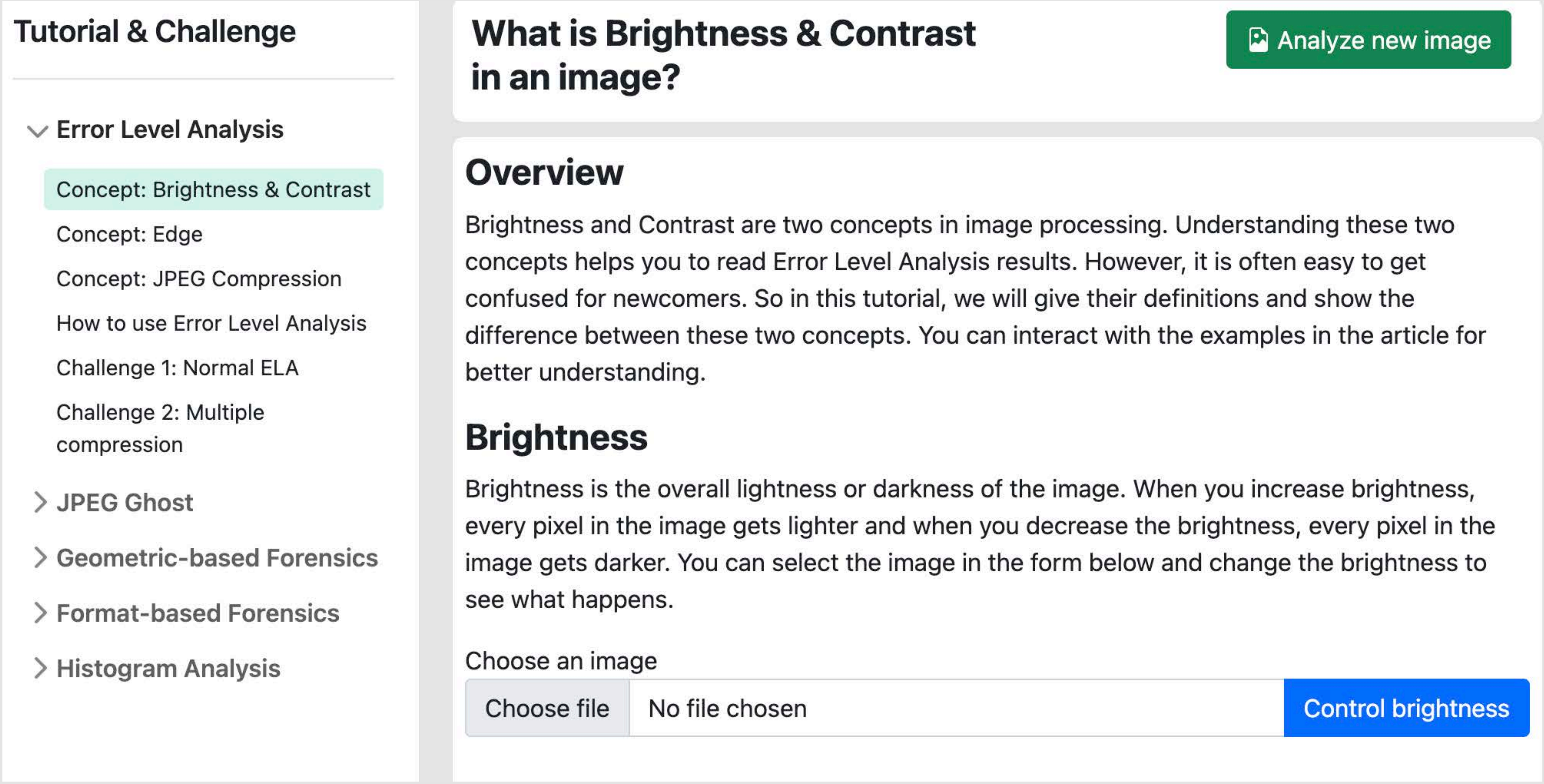
Conditions of using fact-checking tools

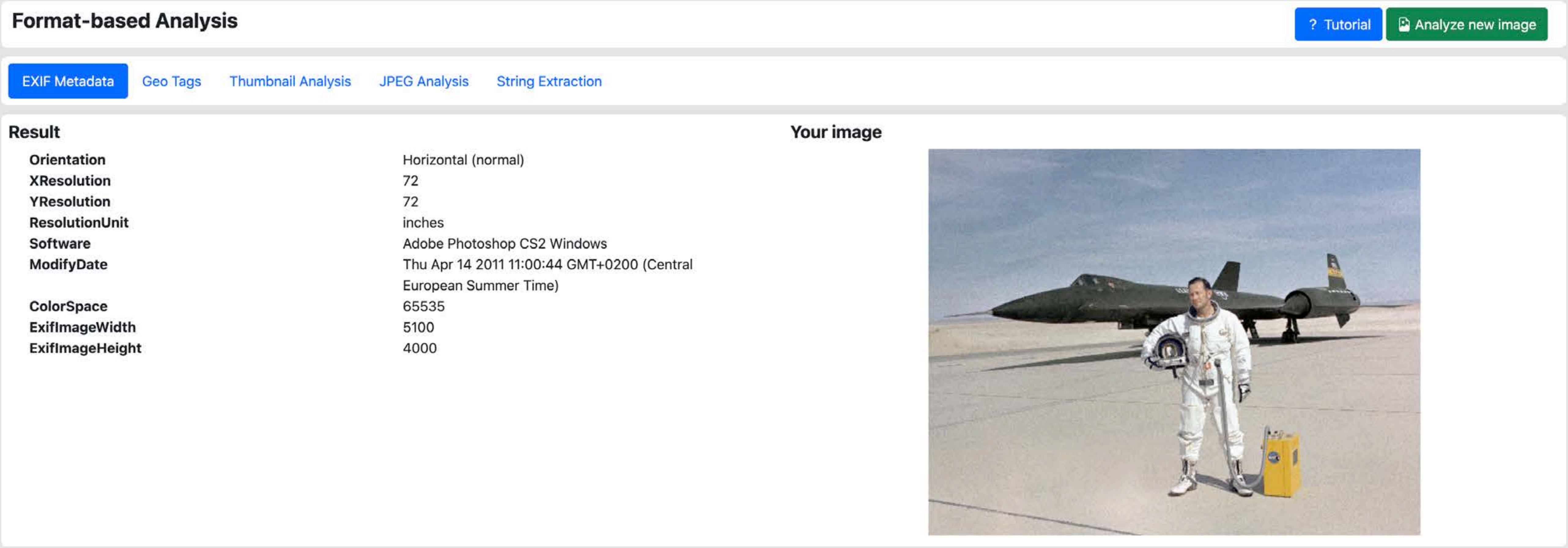
- Tools are not magic wands, they do not define the fact-checker
- Trusting the tools and the results they provide = explicability
- Tools' reliability suppose a shared-expertise between the human and the tool
- Accuracy = considering the context (facts are context-dependent)
- Automation = quality of training datasets + accuracy of the results

"I cannot trust what 's written in this Wikipedia post, because (...) anyone can write that."

JFC8, Sweden

We are developing user friendly, responsible and collaborative Prototypes

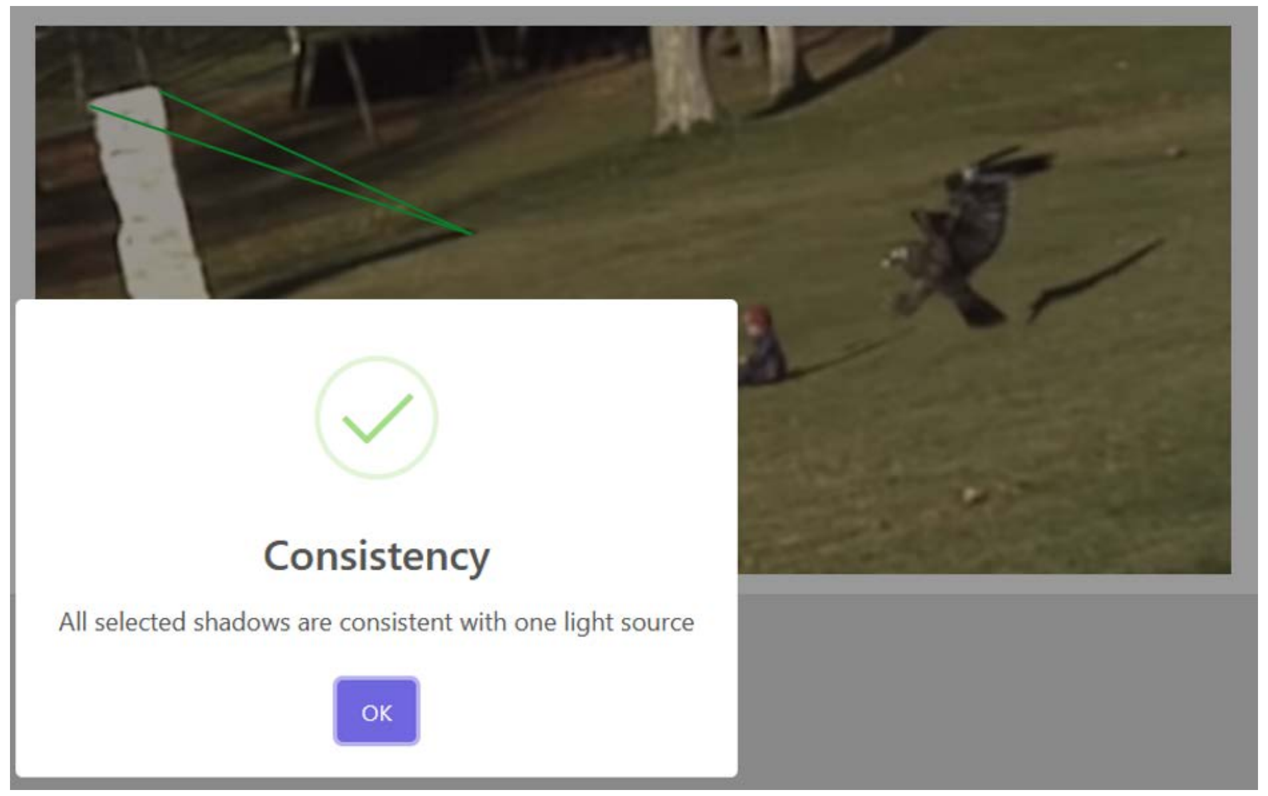


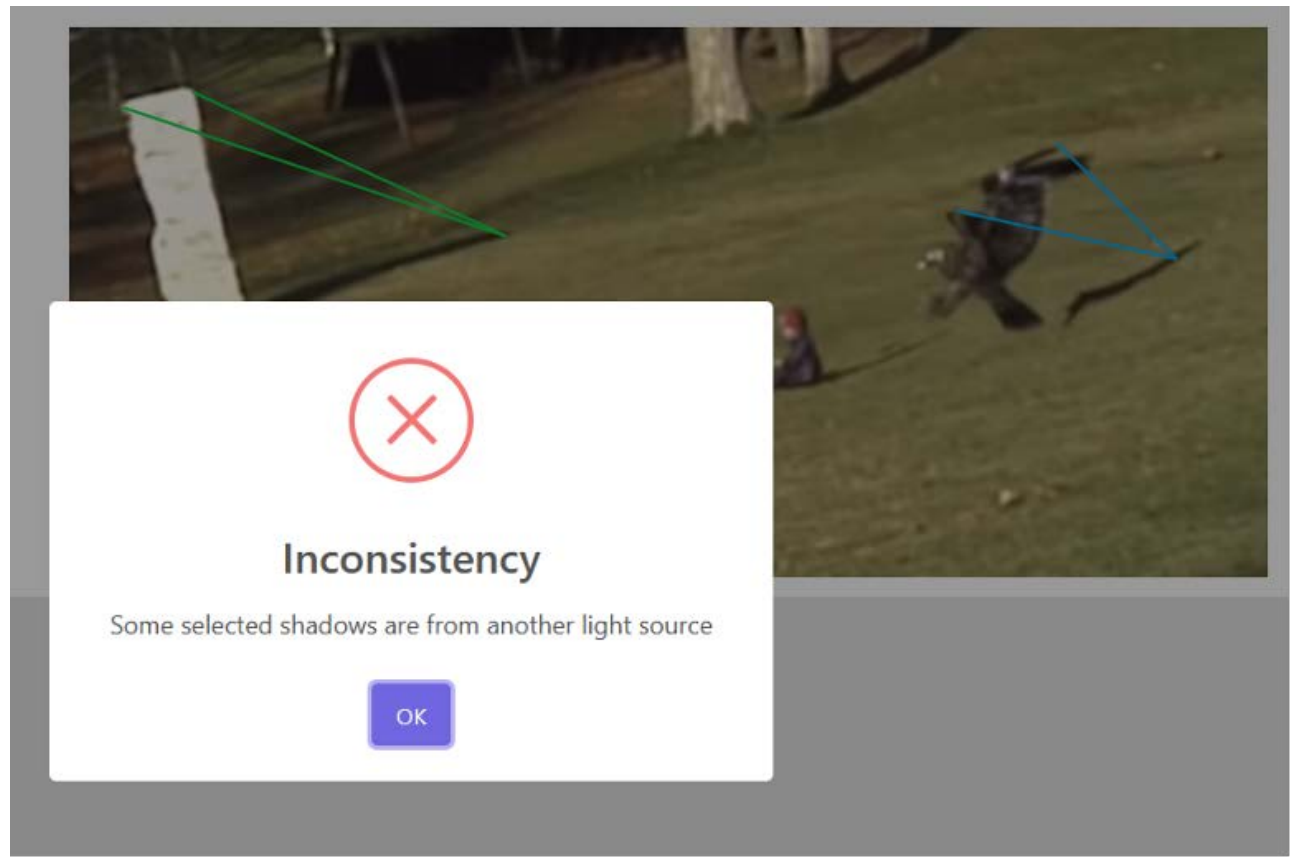


http://dedigi.tech, *"DeDigi: A Privacy-by-Design Platform for Image Forensics"*, Intelligent Cross-Data Analysis and Retrieval, ACM ICMR 2022

Facilitates advanced analysis

Youtube Video:https://www.youtube.com/watch?v=CE0Q904gtMI





http://dedigi.tech, *"DeDigi: A Privacy-by-Design Platform for Image Forensics"*, Intelligent Cross-Data Analysis and Retrieval, ACM ICMR 2022

Summary

- RQs are done very according to the plan
- Outputs:
  - A repository
  - A demo
  - 3 accepted conference papers
  - 1 journal (submitted)
  - 1 research challenge organised

Media Futures

Media Futures

Tusen Takk!

Research Centre for Responsible Media Technology and Innovation
Project number 309339