Abertay University

An examination of reverse image search recognition using deepfaked images

[Redacted]

Introduction to security – CMP110

BSc Ethical Hacking Year 1

2020/21

.

Abstract

This report was an examination of deepfake technology in relation to reverse image searching. It was conducted through taking images of celebrities in deepfaked videos and putting them through various reverse image search engines in order to determine if the engines could correctly identify either the scene (as most deepfakes are from famous films) or the deepfaked actor within the scene. This was done in order to determine if deepfaked content could be used to maliciously alter the results of open source intelligence investigations which frequently use reverse image search engines. Ultimately, it was discovered that it is unlikely that OSINT investigations could be poisoned, as some search engines easily identify scenes from single image frames and therefore could likely find the source image of a given deepfaked image - ensuring that discrepancies could be found and easily proven to be faked.

.

Contents

.

1 INTRODUCTION

1.1 BACKGROUND

To begin with some concepts must be understood. The term “deepfake”, in this report, refers to a piece of video content in which an individual's likeness has been swapped with another individual's likeness through the use of software. This is a broad definition and would include instances in major films and movies where actors' faces have been manipulated through things such as complex video editing and manual face replacement - or even a Snapchat filter, in our case this is less true - the instances looked at here make use of modern machine learning tools available to the public to automate this likeness swapping process.

This is a problematic technology because giving anyone with a sufficiently powerful computer the ability with enough video or images of another person to place them in any other piece of video footage has a huge potential for malicious action on both small and large scale. Whilst there is little to no evidence of a large scale attack using this technology, the small scale malicious actions account for most deepfakes - in 2019 it was estimated 96% of deepfakes were pornographic, with 99% being examples of female celebrities being nonconsensually deepfaked onto porn stars. This problem will only grow, during 9 months it was estimated the number of deepfakes had doubled, a new problem now is that we cannot gauge how many deepfakes there are as many communities making deepfakes went underground as the technology became more prolific and ethical questions were being raised.

Currently, the amount of data (images, video) required by these programs to produce anything of a quality indistinguishable from an unchanged video is the limiting factor, the current programs being used require an immense amount of data being run on very powerful machines for long periods. Time and a powerful computer are easier to obtain than images/video however, which is why the majority of deepfakes are conducted on celebrities - as there is already an abundance and variance of content where their likeness appears. This is only currently true and as the technology advances these barriers will be increasingly lessened.

Within this, one of few ways that the public can accurately identify manipulated images is through the means of reverse image searching, essentially giving an image to a search engine to search for the specific image or similar examples to the image. If the source of an image can be found through this means, discrepancies can be identified between the manipulated image and the one found online - which essentially ensures that it cannot be trusted. This is done using complex algorithms not available to the public, with each search engine having its own specific means of taking an image, identifying what it is and then searching the web for it which in turn means each has its strengths, weaknesses and overall quality.

This is prescient as one the means used to identify individuals in open source investigations is reverse image search, if a reverse image search identifies an individual within a deepfaked image as the person it was intended to be a fake of, it opens the potential for the usage of deepfaked images to be used to maliciously poison open-source investigations.

This technology has some minor issues in that it can be used maliciously to identify people who may not wish to be identified or to get somebody's identity from a single image allowing you to stalk social media or something similar. Currently, the technology available publicly is of varying quality, depending on the search engine and image being used it often fails at identifying simple existing images and cannot provide a source. There is a question of if this is intentional due to the potential risk outlined above, however, due to the lack of open source it is impossible to verify if this is true.

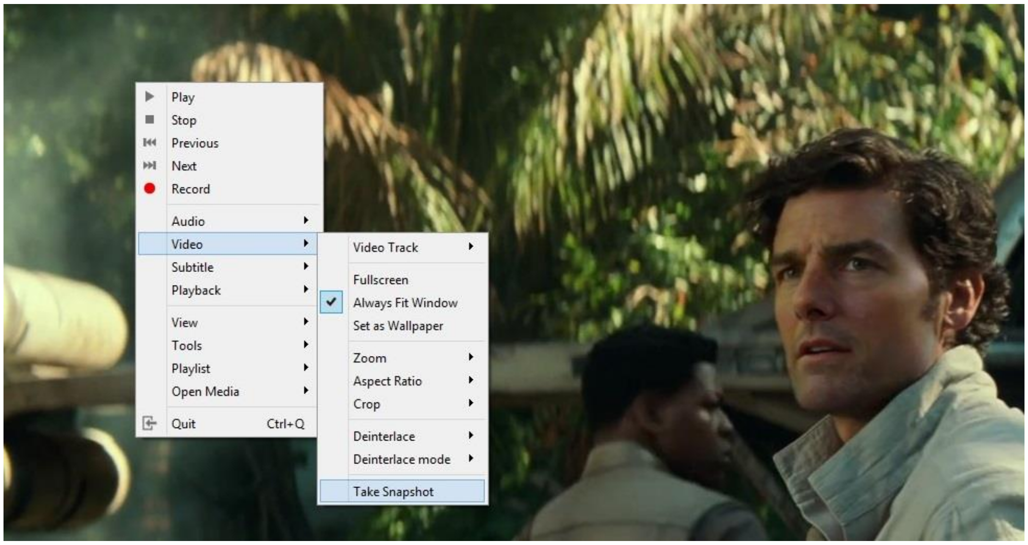
1.2 AIM

The intention of this project is to test reverse image search engines in order to identify any potential deepfakes have been used to manipulate the results. I expect to test a series of high quality deepfaked images taken from deepfaked videos in a series of reverse image search engines, the results of this will then be assessed based on if there is recognition of the scene or recognition of the deepfaked subject.

2 PROCEDURE

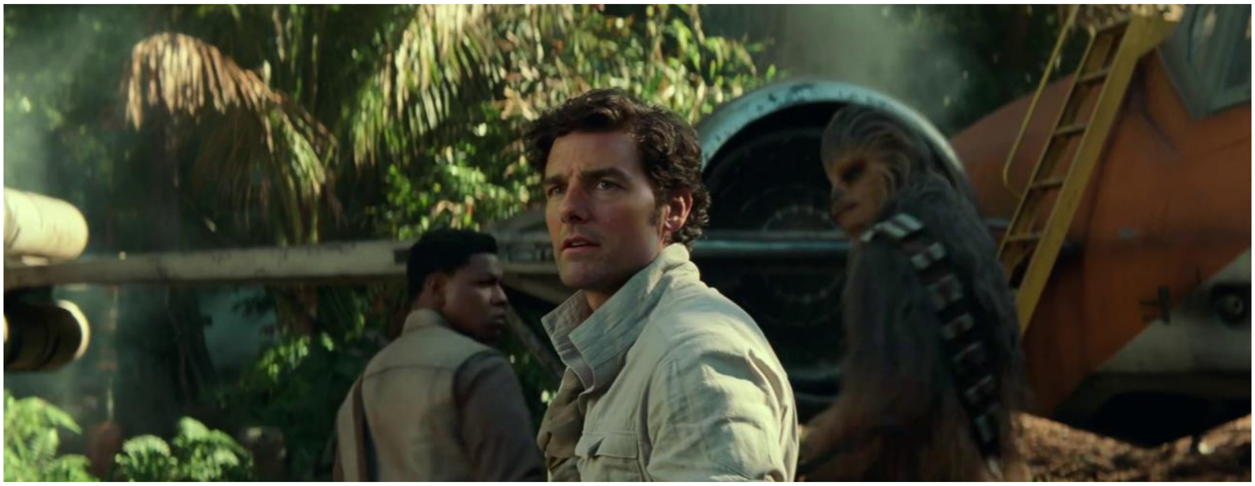
2.1 OVERVIEW OF PROCEDURE

First and foremost to obtain deepfakes of a satisfactory quality a variety of videos were selected from the SFW deepfakes subreddit (Reddit. 2020a), this was used purely as it collated high quality deepfaked content conveniently. Once the videos were selected they were downloaded from Youtube at 1080p, and then placed into VLC media player, where the screenshots were taken. See figure 1..



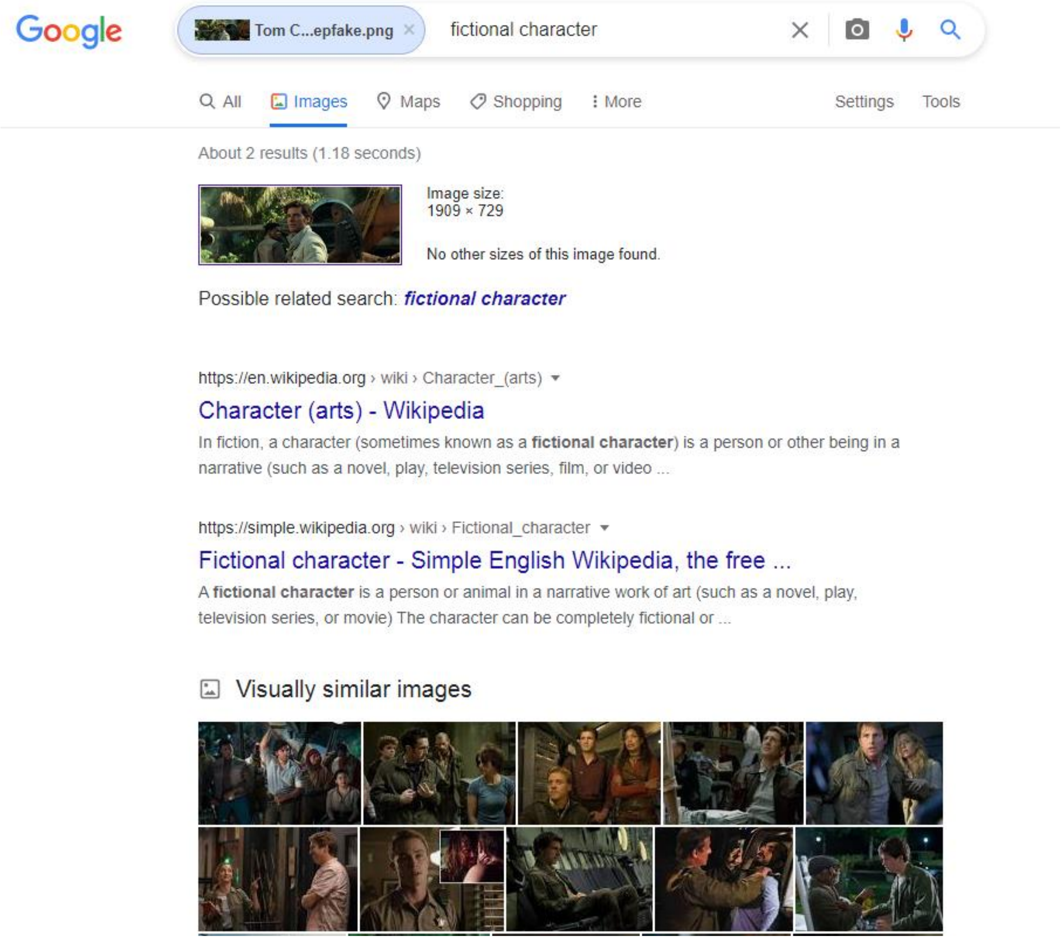
[Figure 1, taking a snapshot of video footage from VLC media player.]

The screenshots were selected from scenes in which the faces of the individuals were undeniably clear in order to give the best chance of detection. Some of the screenshots were cropped to remove watermarks in order to ensure that the only thing being interpreted was the image. See figure 2 for an example of the final images taken.



[figure 2 example of finalized image (Tom cruise deepfaked onto a scene depicting Star Wars character Poe Dameron played by Oscar Issac)]

Once these images had been collected each one was placed into the respective reverse image search engines of Google(Google,2021b), Bing(Bing,2021c) and Yandex(Yandex, 2021d). Based on the results each was reviewed and rated by the researcher based on their recognition of the individual or recognition of the scene. See figure 3 for an example of the results provided when entering an image into Google.



About 2 results (1.18 seconds)

Image size:
1909 × 729

No other sizes of this image found.

Possible related search: ***fictional character***

https://en.wikipedia.org › wiki › Character_(arts)

Character (arts) - Wikipedia

In fiction, a character (sometimes known as a **fictional character**) is a person or other being in a narrative (such as a novel, play, television series, film, or video ...

https://simple.wikipedia.org › wiki › Fictional_character

Fictional character - Simple English Wikipedia, the free ...

A **fictional character** is a person or animal in a narrative work of art (such as a novel, play, television series, or movie) The character can be completely fictional or ...

[Figure 3, a screenshot of results from placing our image into Google reverse image search].

3 RESULTS

[See appendix A for the images reviewed and comments on each image tested.]

Google

Google overwhelmingly disappointed, being ranked last the most. It generally failed to identify the actual image and provided very general descriptions such as “Scene” or “Fictional character” - or even getting it totally wrong in one instance, identifying the type of character incorrectly which was highly misleading. Generally the “Visually similar images” was the most useful, and did provide some images of characters or actors from the same film, however it only twice provided images from the same scene or the same frame. Meaning unless already aware of the actor behind the deepfake or the scene the film was from, it is unlikely you would be able to pick these images out from the variety of others. It never identified the deepfaked actor and this could be pinned down to it’s overwhelmingly poor performance overall.

Bing

Bing had an interesting result, in that it was exceptionally good at interpreting the faces within an image. It identified the deepfaked actor every single time. In terms of the aim, this means it is an abject failure in that the deepfaked face fooled Bing every time. Out of the 3, it still does come second though - as it did loosely identify some scenes but never provided the same frame, and it’s ability to be tricked is actually somewhat of a useful feature when combined with other reverse image search engines. It furthermore did impress with its successful identification of one of the source videos for the image despite the difference in thumbnail.

Yandex

Yandex was overwhelmingly superior to the other two, it frequently provided the original frame from a given deepfaked image and successfully identified the film and original actor within the frame. It failed to be tricked by the deepfaked images every time. In addition, it provided links to examples of the frame on the internet - making tracking a given source image down much easier to find discrepancies. It is exceptional for usage of identifying when an image has a facial replacement.

3.1 GENERAL DISCUSSION

One of the more interesting things observed is that Yandex is exceptionally good at recognizing individual frames from films, however is essentially cheating. Based on observation it identifies the frame, then simply finds terms related to the images it’s found, and it’s also very good at this - this is why despite the lack of the actor Hugh Jackman in the deepfake, it identified him - because it identified the frame then pulled information based upon that - it isn’t actually doing any recognizing from the faces within the image because fundamentally he wasn’t in it. This is in direct contrast to Bing, which seemingly overwhelmingly prioritises faces and it is very good at identifying them from an image.

With a combination of these two reverse image search engines, it would be likely possible to identify the source of a deepfaked image, the individual within it and the individual intended to be the deepfaked.

3.2 Conclusions

- Yandex is overwhelmingly the preferred engine for identifying deepfaked content source images, and therefore the most useful for preventing the malicious use of them in OSINT investigations
- Bing is exceptionally good at identifying faces, which while useful, is fooled by deepfaked images and identifies them as the deepfaked individual - therefore making it the most likely to allow the malicious use of deepfakes. When used in conjunction with other reverse image search engines however, it becomes a useful tool for identifying a hypothetically deepfaked person who may not be aware of it.
- Google is essentially not useful for this purpose due to it's overwhelmingly bad performance at identifying anything.

3.3 Future Work

In an ideal world, obtaining the source code for the various image recognition algorithms would give an immensely more clear answer as to why the results obtained were what was found, however this is very unlikely. More practically testing a wider variance of deepfaked images, such as those of objectively lower quality would give a clearer picture as to the overall quality of these search engines for usage in open source investigating.

1. References

Ajder, H. Patrini, G. Cavalli, F. Cullen, L. 2019. *The State of Deepfakes: Landscape, Threats, and Impact*. [Report] Available from: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf [Accessed 13th April 2020]

Reddit. 2021a. *Deepfakes that are Safe for Work.* [online]. Available from: https://www.reddit.com/r/SFWdeepfakes/ [Accessed 13th April 2020]

Google. 2021b. Reverse image search. [online]. Available from: https://www.google.com/imghp?hl=en [Accessed 13th April 2020]

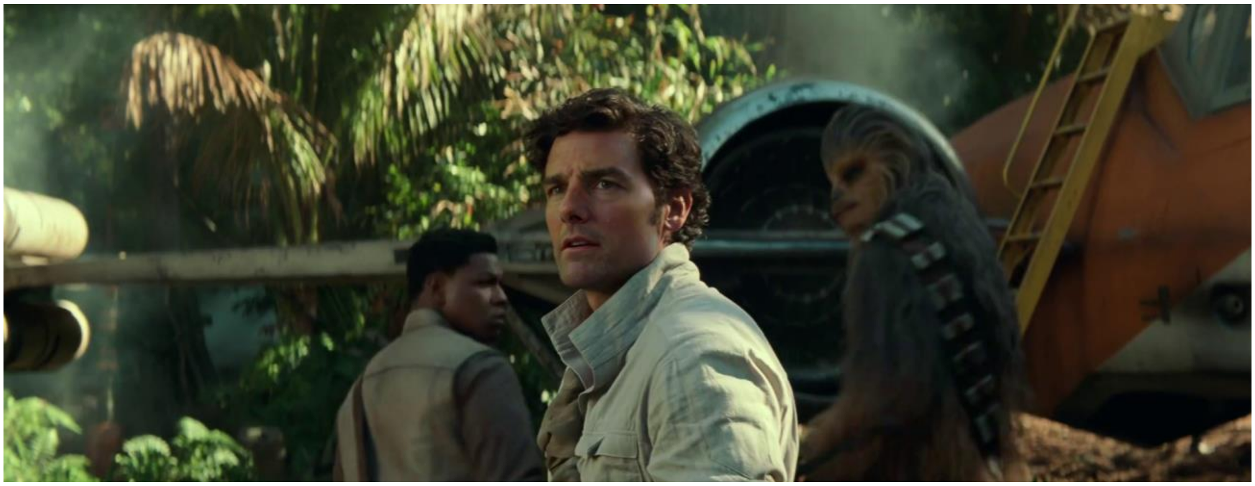
Bing. 2021c. Visual search. [online]. Available from: https://www.bing.com/visualsearch?FORM=ILPVIS [Accessed 13th April 2020]

Yandex. 2021d. Reverse image search. [online]. Available from: https://yandex.com/images/ [Accessed 13th April 2020]

2. APPENDICES

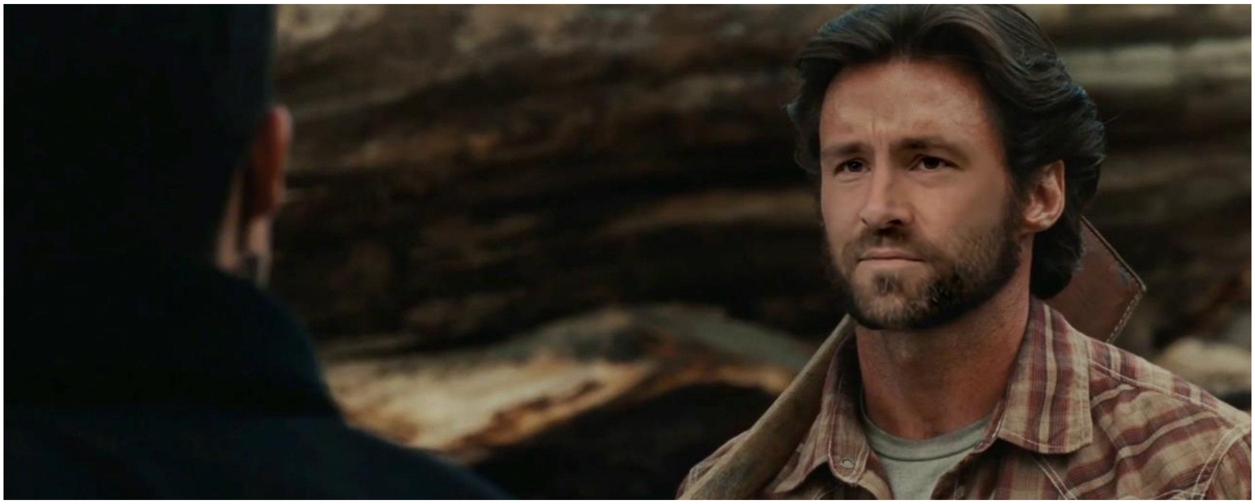
1. APPENDIX A

Image 1: Tom Cruise deepfaked onto the character Poe Dameron, played by Oscar Issac.



Engine	Comments	Rank
Google	Failed to identify the actor, simply recognising it was a “Fictional character.” Under visually similar images it also failed to identify the scene - interestingly it did have both instances of Tom Cruise and Poe Dameron, but it was amongst a variety of other unrelated film scenes, therefore doesn’t count.	3
Bing	Successfully identified the actor as Tom Cruise, absolutely failed to identify the scene.	2
Yandex	Correctly identified the scene from the film to the degree of giving multiple examples of the original frame, the name of the film and name of the character. Did not identify Tom Cruise even slightly.	1

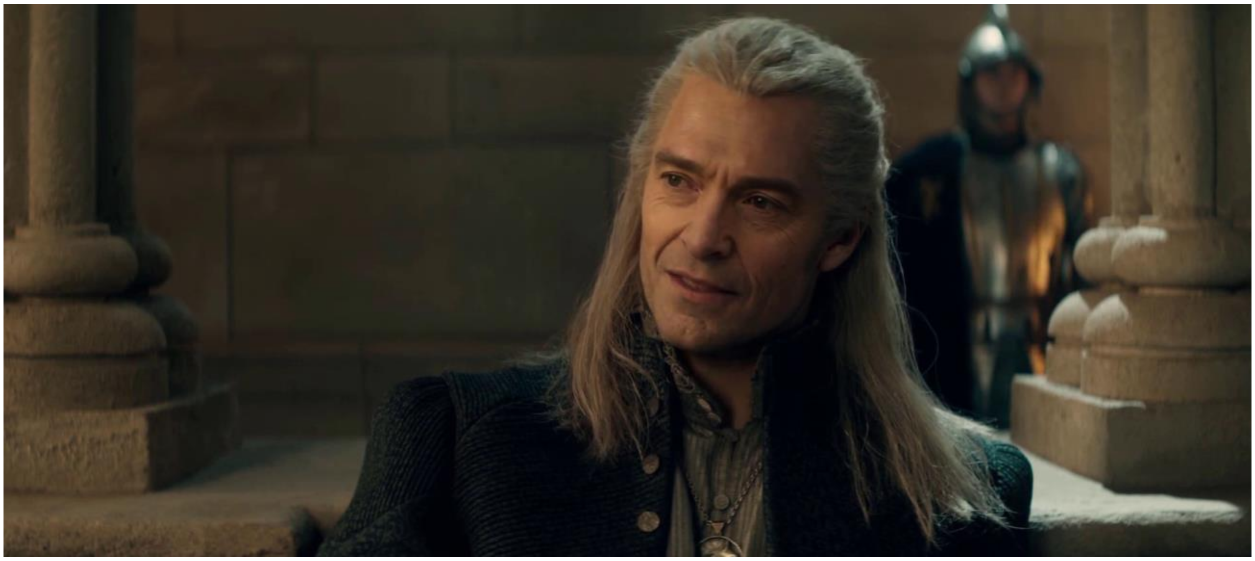
Image 1: Ryan Reynolds deepfaked onto the character Wolverine, played by Hugh Jackman.



Engine	Comments	Rank
Google	Absolute failure to identify the scene, simply giving the result that it was a “Scene”. Furthermore additional failure to identify either Ryan Reynolds or Hugh Jackman. Similar images did provide one of Hugh Jackman as Wolverine, however it is from a film in which he is not similar to the above image and amongst a variety of other images that were not visually similar, therefore counting as a failure.	3
Bing	Correctly identified Ryan Reynolds as the deepfaked individual. Absolute failure to identify the scene, seemingly did not even try, all of the related content was about Ryan Reynolds.	2
Yandex	Correctly identified the scene, providing examples of frames the scene was from and even the year the film came out. Identified Hugh Jackman despite his lack of appearance. Did not identify Ryan Reynolds at all.	1

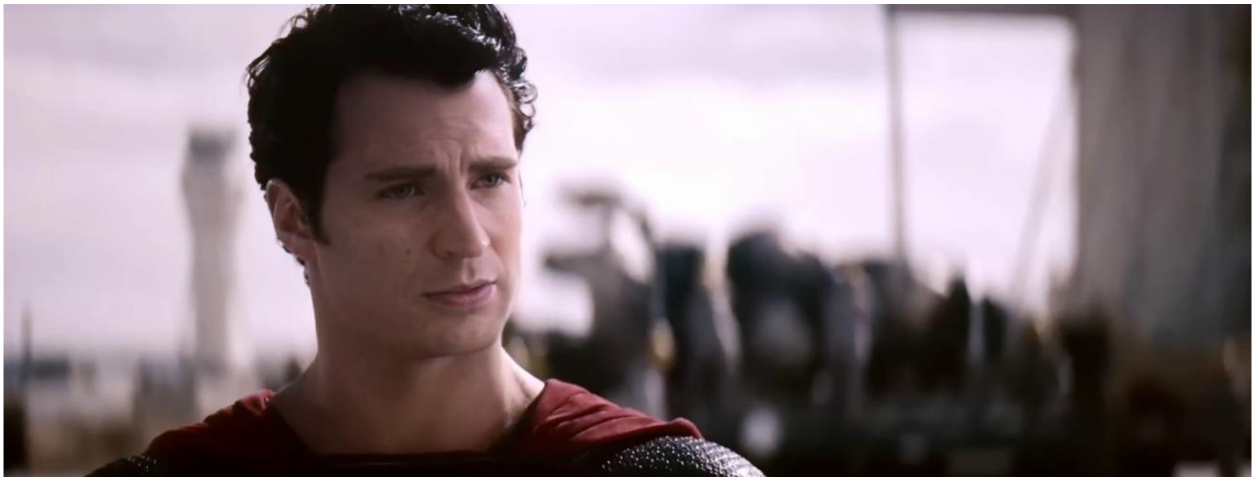
Image 3:

Hugh Jackman deepfaked on to the character Geralt of Rivia, played by Henry Cavill



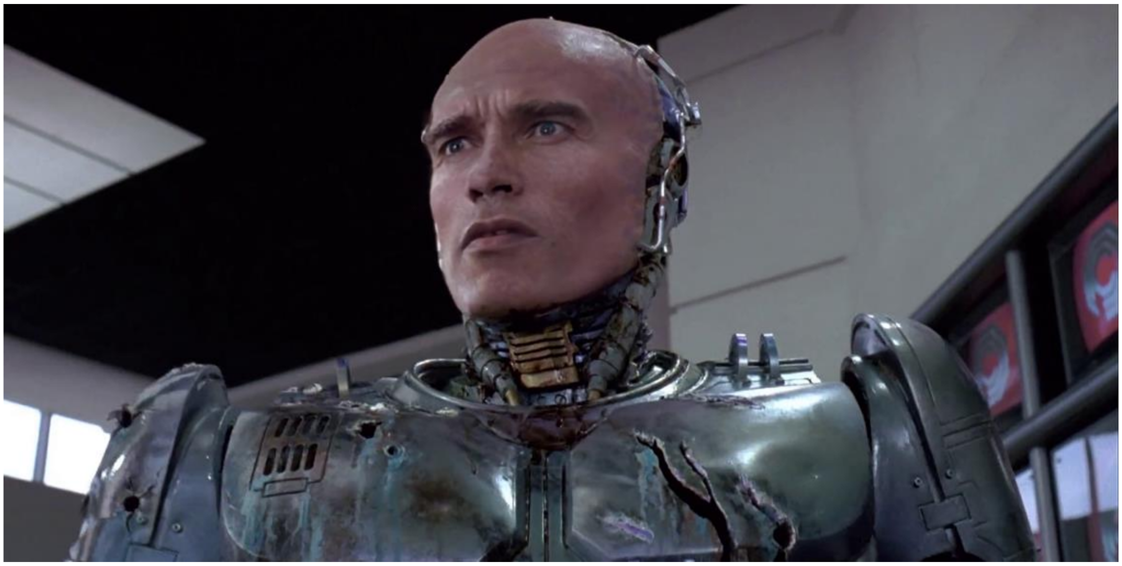
Engine	Comments	Rank
Google	Correctly identified the scene, providing frames from the same scene in visually similar images. Did not specifically identify Henry Cavill, simply identifying a “Fictional character.” Did not identify Hugh Jackman at all	3
Bing	Correctly identified the original scene impressively did not identify Henry Cavill as the individual Correctly identified hugh Jackman as the individual in the image Additionally: Very impressively managed to provide the video the image had been sourced from, despite cropping it and the thumbnail being significantly different, only containing a similar frame for half of it. Evidence suggests it may have identified the scene and individual and managed to put them together to find the original video.	1
Yandex	Correctly identified the original scene, impressively even mentioned the content of the dialogue involved in the scene as well as the series and character. Did not identify Henry Cavill specifically. Did not identify Hugh Jackman at all.	2

Image 4: Chris Evans deepfaked onto Superman, played by Henry Cavill



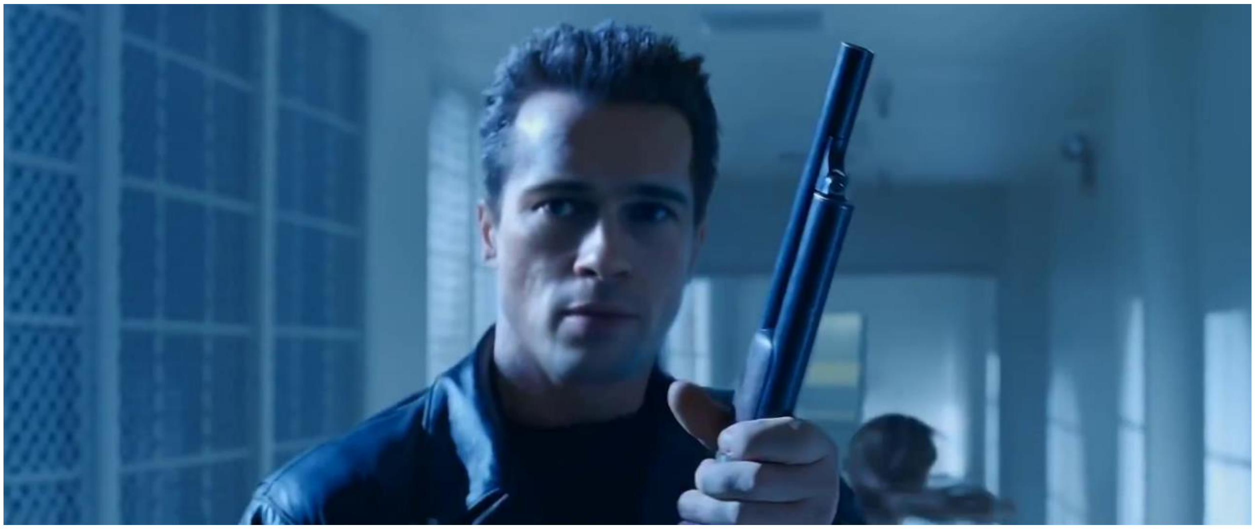
Engine	Comments	Rank
Google	Failed to identify the original scene, but correctly identified the character and several frames from the same film. Failed to identify Chris evans completely. Identified Henry Cavill as the actor in the original scene	3
Bing	Failed to identify the original scene Correctly identified Chris Evans within the deepfaked image Did not identify Henry Cavill as the original actor	2
Yandex	Correctly identified the original scene and character Failed to identify Chris Evans Completely Identified Henry Cavill as the actor in the original scene	1

Image 5: Arnold Schwarzenegger deepfaked onto Robocop, played by Peter Weller.



Engine	Comments	Rank
Google	Correctly identified the scene, film and year. Failed to identify Arnold Schwarzenegger completely failed to identify Peter Weller as the original actor	3
Bing	Correctly identified the film Correctly identified Arnold Schwarzenegger failed to identify Peter Weller as the original actor	1
Yandex	Correctly identified the scene, character and film. Failed to identify Brad Pitt completely Correctly identified Arnold Schwarzenegger as the original actor	2

Image 6: Brad Pitt deepfaked onto The Terminator, played by Arnold Schwarzenegger.



Engine	Comments	Rank
Google	Incorrectly identified the film, suggesting "Star Wars Characters" Failed to identify Brad Pitt Failed to Identify Arnold Schwarzenegger	3
Bing	Failed to identity the film Successfully identified Brad Pitt Failed to Identify Arnold Schwarzenegger	2
Yandex	Correctly identified the scene, character and film, even giving the year. Failed to identify Brad Pitt Correctly identified Arnold Schwarzenegger as the original actor	1