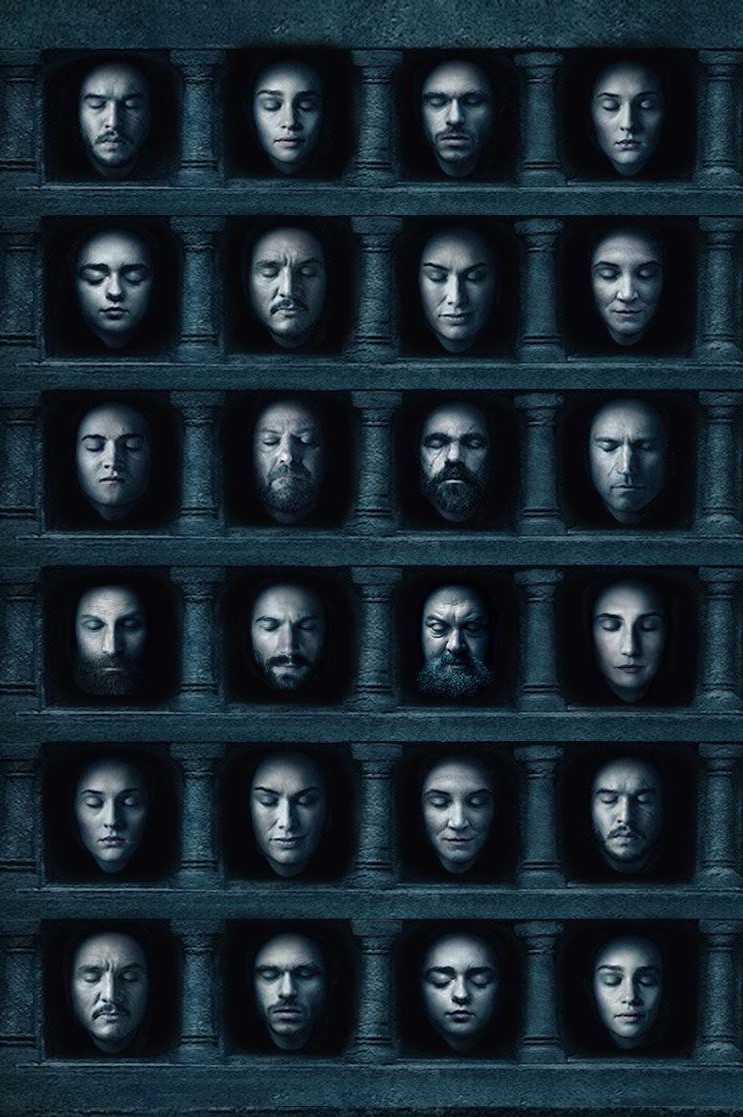
|  |
| --- |
| **SENTINEL®:**  Safeguarding Your  Online Honor  Computer Vision Project |
|  |



### **Executive Summary**

The last two decades have seen a tremendous increase in the development of AI capabilities. Key developments in hardware such as faster processors capable of parallel computing as well as software that can leverage these developments to process data faster, combined with increased interest in fields such as deep neural architectures have led to numerous applications. These developments have led to productivity improvement tools and lifesaving applications such as cancer detectors from scanned biopsy images. But, just like any other technology, deep learning can be misused for malevolent purposes.

A common form of this misuse is AI-generated synthetic media that have seen an increase in malevolent use commonly called as Deepfakes:

Deepfakes (a portmanteau of "deep learning" and "fake") are synthetic media in which a person in an existing image or video is replaced with someone else's likeness. While the act of faking content is not new, Deepfakes leverage powerful techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content with a high potential to deceive…

…Deepfakes have garnered widespread attention for their uses in celebrity pornographic videos, fake news, hoaxes, and financial fraud. This has elicited responses from both industry and government to detect and limit their use.

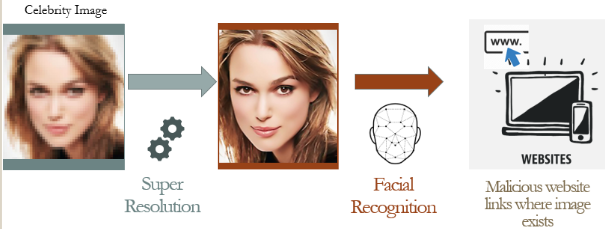
Below is an example of how a deepfake has been used to morph the face of actor Steve Buscemi on the actual image of Jennifer Lawrence (**Click on the play button**):

We propose the use of deep learning techniques to create an app “Sentinel”, where users can get a report of websites where malicious content with their faces was seen so that they can be removed. Sentinel uses facial recognition and super-resolution models to detect undesirable media on platforms known to host such malicious content such as photo and gif buckets (Imgur, Gfycat), pornographic websites, and those on the dark web. Sentinel provides users a web/mobile interface where an individual can upload their picture and see if there is any content that matches these known malicious platforms. The target content is matched using facial recognition on deep faked images and video thumbnails extracted from a list of websites known to host malicious content.

The application is in its initial stages and cannot yet differentiate between non-deep fake content and deep fake content. Moreover, the application will not be able to identify malicious content if the thumbnails do not contain an image of the person. It is assumed that all content on websites identified as malicious is undesirable and needs to be taken down.

### **Methodology**

1. User uploads their pictures (preferably taken at various angles) to Sentinel.com,
2. Sentinel.com scrapes thumbnails from malicious websites such as pornographic websites, image/gif buckets (Imgur, Gfycat), etc.
3. Thumbnail images are enhanced using a super-resolution network and are used as test images.
4. Images uploaded by users are treated as a training dataset and the VGGFace model is used to identify if the thumbnail contains the user’s face.
5. Sentinel provides a report with links to malicious content, which can then be reported to authorities to be taken down.



### **Assumptions**

1. All content on malicious websites is undesirable – even if it is not a deep fake.
2. Users will provide adequate photos to Sentinel.com from various angles

### **Results**

Cameron Diaz, a popular actress was chosen as a fictional client. The VGGFace model was trained using the below anchor images of Cameron:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AA person smiling for the camera  Description automatically generated** | **BA person smiling for the camera  Description automatically generated** | **CA person smiling for the camera  Description automatically generated** | **DA person smiling for the camera  Description automatically generated** | **EA person smiling for the camera  Description automatically generated** |

The test set included a mix of valid test images (Resolution enhanced - Thumbnail screenshots of deep fakes from malicious websites) as well as images of other actresses, to see how well the model could differentiate between positive and negatives cases:

|  |  |  |  |
| --- | --- | --- | --- |
| **1A close up of a person  Description automatically generated** | **2A person posing for the camera  Description automatically generated** | **3A picture containing wearing, shirt  Description automatically generated** | **4A person taking a selfie  Description automatically generated** |
| **5**  **A close up of a person who is smiling at the camera  Description automatically generated** | **6**  **A person smiling for the camera  Description automatically generated** | **7**  **A person sitting on a table  Description automatically generated** | **8**  **A person posing for the camera  Description automatically generated** |
|  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Distance Metric (Cutoff = 0.9, Green = Match)** | | | | | | | |
|  | | **Test Images** | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **T**  **R**  **A**  **I**  **N** | **A** | 0.83 | 0.91 | 0.96 | 0.84 | 0.99 | 0.53 | 1.23 | 0.96 |
| **B** | 0.83 | 0.81 | 1.03 | 0.85 | 1 | 0.47 | 1.22 | 0.84 |
| **C** | 0.90 | 0.89 | 1.06 | 0.95 | 1.07 | 0.54 | 1.29 | 0.87 |
| **D** | 0.80 | 0.86 | 0.98 | 0.81 | 0.92 | 0.43 | 1.20 | 0.94 |
| **E** | 0.80 | 0.92 | 1.05 | 0.88 | 1.08 | 0.40 | 1.26 | 0.82 |

Using a cutoff score of 0.9, we have image **1,2,4,6,8** as matches. Images 1,2,3 & 4 were thumbnails clipped from malicious websites and are deepfakes. Sentinel was unable to match image #3, because of the odd angle at which Cameron is resting, and possibly because of the overexposed nature of the image – a training image that captures this face angle could help in matching this image.

Images 5,7 & 8 were intentionally kept as non-matches and are images of Angelina Jolie, Jennifer Aniston, and Ellen Barkin, respectively. This was done to understand what an adequate cutoff score would be. It was interesting to note that the algorithm showed a significant score for Ellen Barkin, an actress who is popularly known to resemble Cameron Diaz.

### **Future Scope and Clientele**

Future iterations of Sentinel will see the following enhancements:

1. Improved ability to differentiate deepfakes from valid content on generic websites such as YouTube. Instagram, Facebook using Generative Adversarial Networks.
2. Improved handling of pose and lighting invariance to handle non-frontal training images

With the explosion in internet usage, Sentinel will become an important tool for every individual and not just those in public life such as Celebrities, Politicians, etc.

Videos are important admissible evidence in courts and Sentinel will play an important role in vetting these videos as original content. In the future, Sentinel is likely to be as important and ubiquitous as the Antivirus for every individual.