YDSP Young Defence Scientists Programme

DSTA Defence Science & Technology Agency

DSO National Laboratories

FOOLING FACIAL RECOGNITION SYSTEMS AND MITIGATION

Members:
Eunice Koh Kexin, Fu Wentao (Claire), Katelyn Kang Jia Xuan (Raffles Girls' School)

Mentor:
Shen Bingquan (DSO National Laboratories)

INTRODUCTION

Deep Neural Networks have been widely utilised in various domains and are susceptible to adversarial attacks. In this research, we present a novel **adversarial patch attack framework** involving differentiable rendering and simulated annealing, and evaluated the effectiveness of various defence measures against it, highlighting the need for more robust defences against such attacks.

PATCH PLACEMENT LOCATION

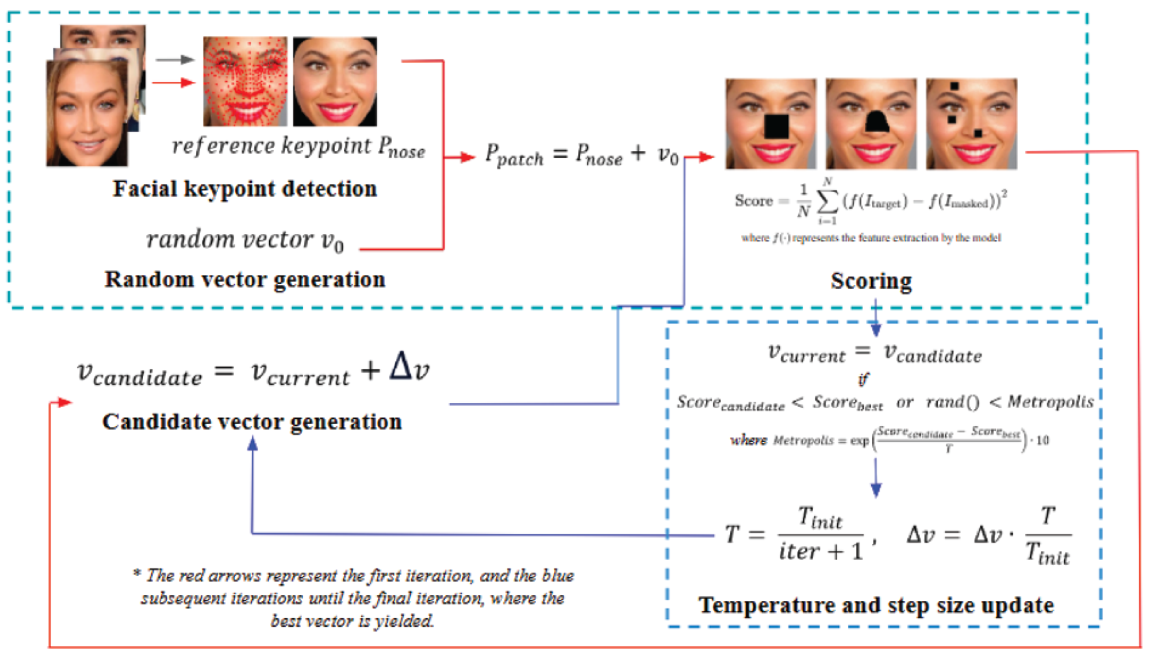


Fig. 1, showcasing Simulated Annealing Pipeline

HEATMAP

This is done by **using a black patch to occlude all possible locations** - features highlighted towards the red end of the spectrum are more salient.

SIMULATED ANNEALING (SA)

Instead of force-fitting the patch at all locations, SA **utilises temperature**. At **higher temperatures**, it is more likely to **accept a worse solution**, with a higher loss, to **avoid being stuck at a local minimum**. The temperature starts high and gradually decreases, eventually converging to an optimal location.

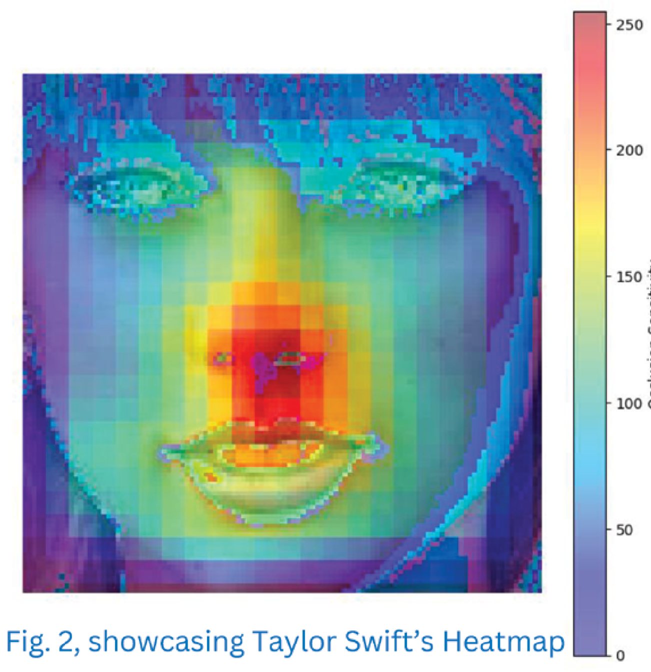
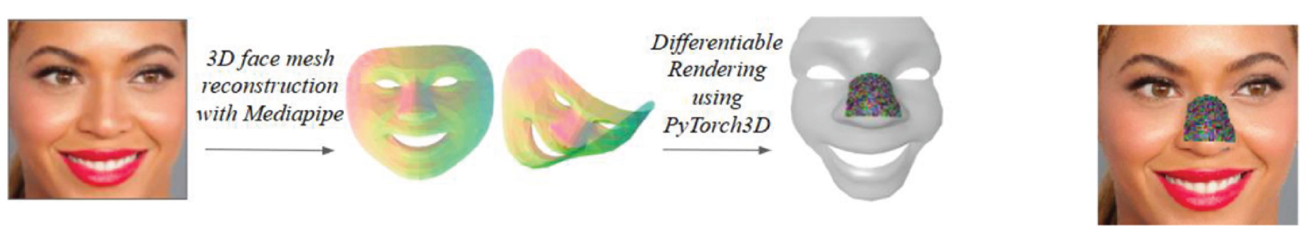


Fig. 2, showcasing Taylor Swift's Heatmap

3D PATCH ATTACK

3D DIFFERENTIABLE RENDERING



We used MediaPipe's face mesh function to **find the predefined face mesh vertices**. Then, using PyTorch3D, the patch was **applied as a texture** mapped to the face mesh, **and rendered** using its differentiable rendering function. This allows for the patch to **simulate real-world conditions**, where factors such as lighting and slight rotations exist.

THIN PLATE SPLINE (TPS)

The patch is constrained at **certain points**, causing it to **bend to a specific shape**. After detecting points via MediaPipe, we bend and warp the patch using **minimum bending energy**. This allows for the patch to **morph based on the attacker's face**.

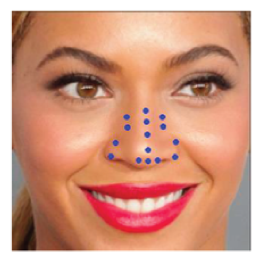


Fig 5.

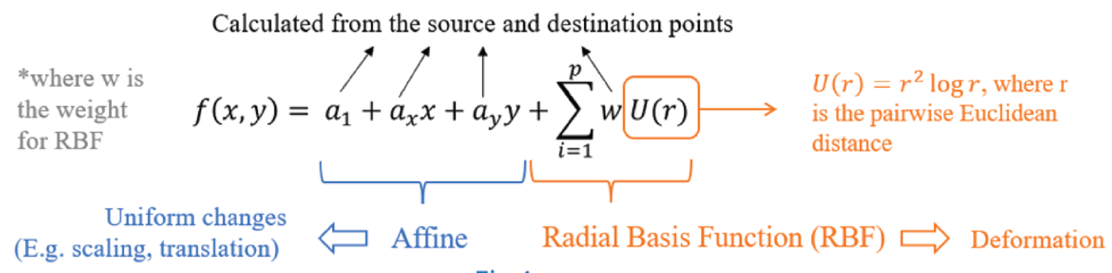


Fig 4.

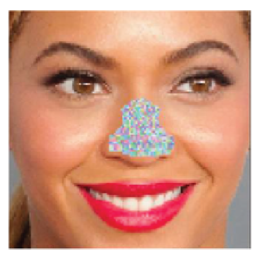


Fig 6.

DEFENCE & MITIGATION

Adversarial patches generated often have **higher frequencies** compared to the rest of the image [2] as patch generation processes rely on some form of **iterative noise** in the patch region [2]. We **extract the frequencies of the image using Fast Fourier Transform (FFT), covered high-frequency regions**, and reconstruct the modified image.

To prevent defensive models from exploiting this loophole, we can utilise FFT in the creation of the adversarial attacks too! This can be done by **extracting the low-frequency components** of the adversarial patch and **optimising it.**

When we tested this new patch, the model was unable to detect it.

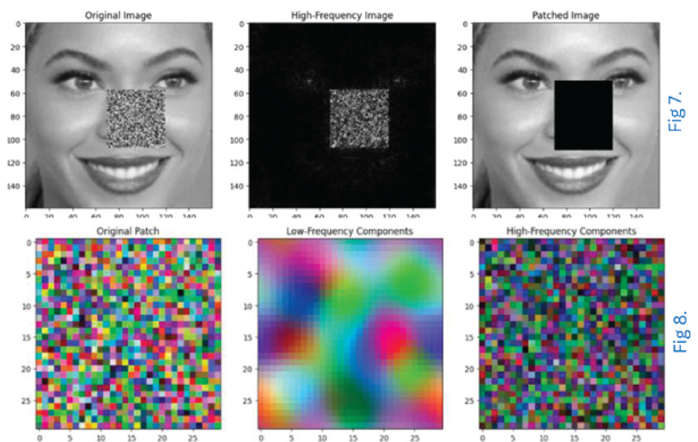


Fig 7.
Fig 8.

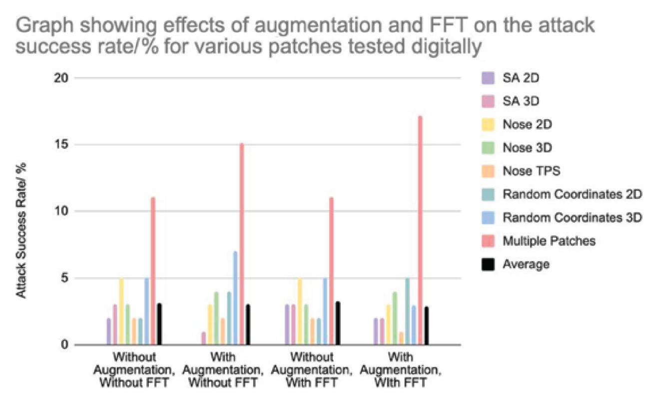


Fig 9.

RESULTS & DISCUSSION

LOCATION

SA was shown to be more effective (74.8% misclassification) than the heatmap (72.5% misclassification) using MSE Loss. Using the defence model, nose coordinates yielded best results likely due to overfitting of an invariant feature.

2D & 3D

Contrary to our predictions, **3D-optimised patches were not much better** than 2D. This could be because the **material** chosen (paper) **did not fold to the features** of the face well and the patch did not bend as much, unlike as assumed in the 3D patches.

MULTIPLE PATCHES

Though **multiple patches** produced a higher MSE loss as compared to single patches, it **was able to fool our defence model better**. This could be due to smaller attack area and lower possible pixel combination, and it could have exploited the defence model's tolerance for small variations.

When testing **in real life, most were unable to meet the threshold for a misclassification**, though there was a decrease in the loss. This could be due to a plethora of reasons: **the camera capturing process, colour aberrations in printing, and a lack of pixel space in real life, which the model was trained on.**

CONCLUSION

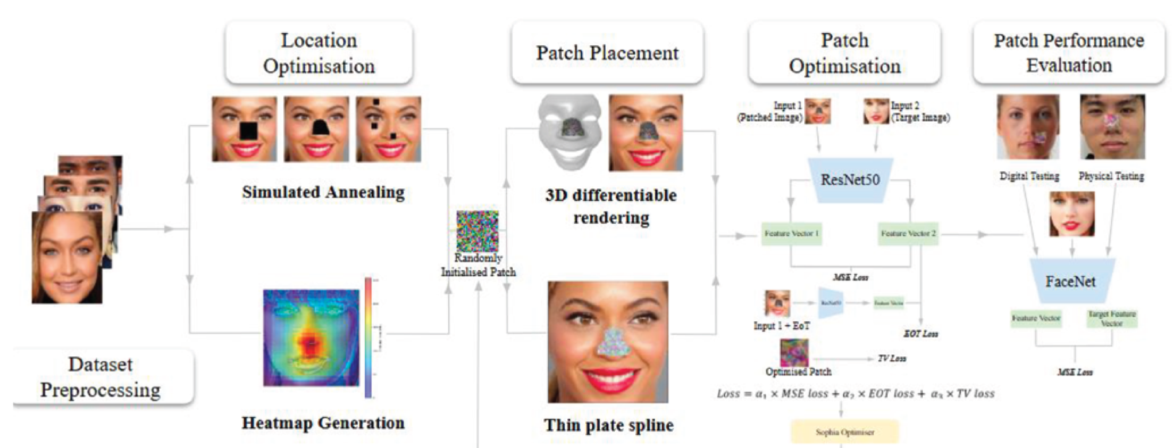


Fig 10. Overall Adversarial Attack Pipeline

Future Work

- Accounting for material properties during 3D differentiable rendering could improve real-world transferability.
- Expand our approach to multiple patches, particularly in exploring patch interaction—sizes, locations, combinations.
- Integrating more sophisticated frequency domain extraction methods, like multi-scale frequency analysis during training, and introducing other parameters to ensure robustness against low-frequency patch attacks.

References:
[1] Mahmood Sharif et al. (2019) A General Framework for Adversarial Examples with Objectives. https://doi.org/10.1145/3317611
[2] Zhun Zhang et al. (2024) Towards a Novel Perspective on Adversarial Examples Driven by Frequency https://arxiv.org/pdf/2404.10202v1

Image of Beyonce taken from: https://www.britannica.com/biography/Beyonce
Image of Taylor Swift taken from: https://i.pinimg.com/564x/38/9c/83/389c8345541206c757ee9d25a910d97b.jpg