Deepfake Detection

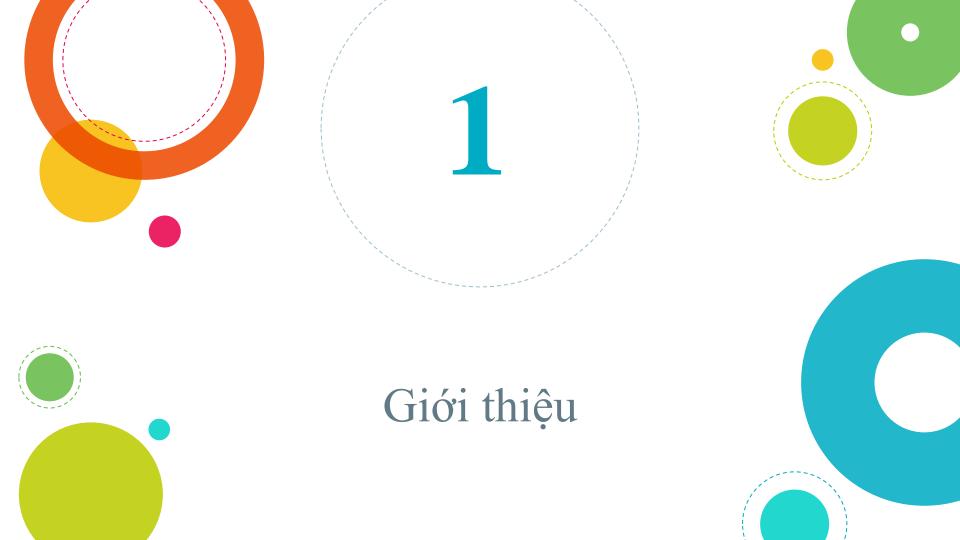


Thành viên:

Họ và tên	MSSV
Nguyễn Vũ Dương	20520465
Lê Trần Quốc Khánh	20520574

GVHD: Mai Tiến Dũng





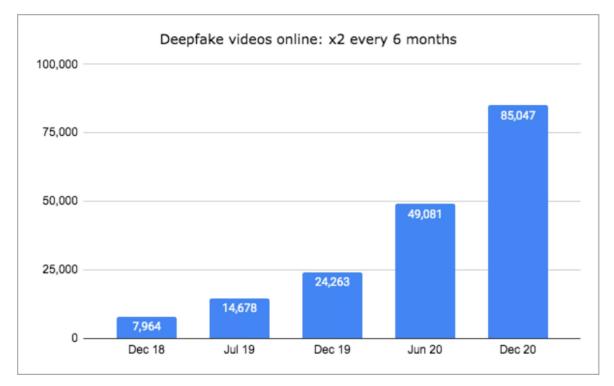
Deepfake là gì?

Deepfake (một từ ghép của "deep learning" và "fake") là một kỹ thuật để tổng hợp hình ảnh con người dựa trên trí tuệ nhân tạo(AI).

Nó được sử dụng để kết hợp và chồng các hình ảnh và video hiện có lên các hình ảnh hoặc video nguồn bằng cách sử dụng một kỹ thuật học máy (machine learning) được gọi là Generative Adversarial Network (tạm dịch là mạng chống đối tạo sinh)



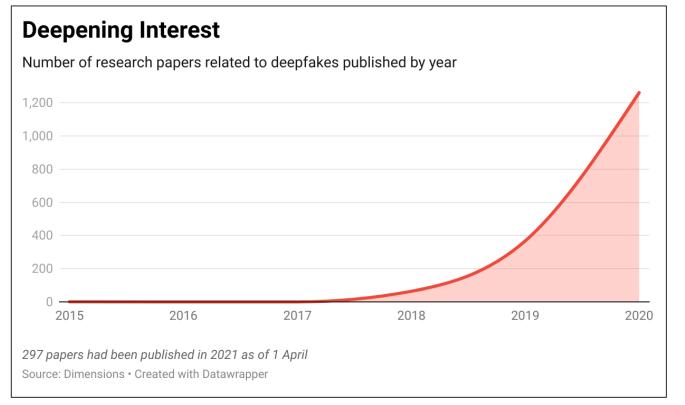
Sự phát triển của các video Deepfake



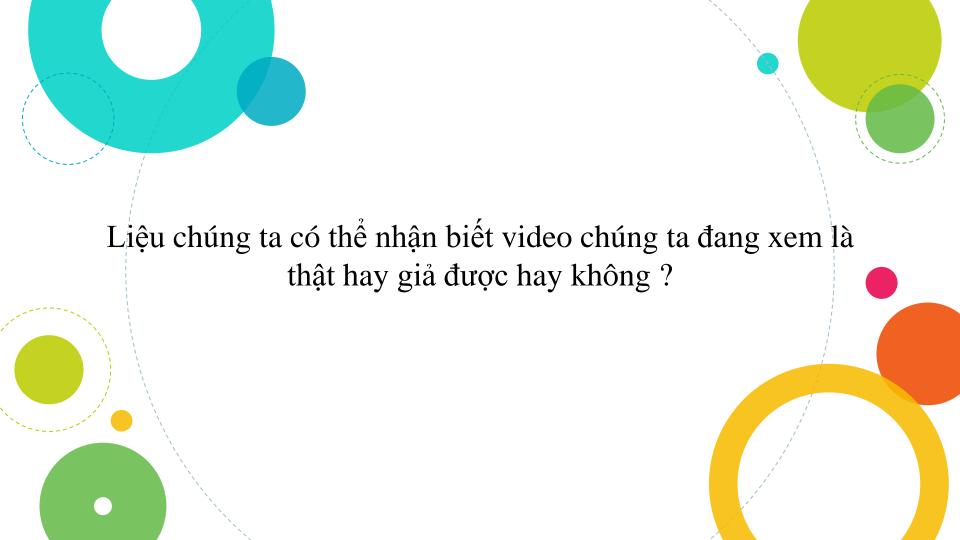
Source: Sensity

Biểu đồ 1: Sự phát triển của video Deepfake từ 12/2018 đến 12/2020

Sự phát triển của các bài báo Deepfake



Biểu đồ 1: Sự phát triển của các bài báo liên quan đến Deepfake từ năm 2015 đến 2020







Thật

Thật hay Giả?

Giả

Giả







Tại sao ta cần phải có công cụ kiểm tra deepfake?

- Tin giả
- Tuyên truyền sai lệch về chính trị
- Bôi nhọ bằng phim khiêu dâm
- Gian lận tài chính (dùng hình ảnh của người ảnh hưởng)









- [1] Artists create Zuckerberg 'deepfake' video https://www.youtube.com/watch?v=cnUd0TpuoXI
- [2] https://www.newsweek.com/trump-deepfake-russia-rutube-1687647
- [3] You Won't Believe What Obama Says In This Video https://www.youtube.com/watch?v=cQ54GDm1eL0



O Công cụ tạo deepfake

- Faceswap
- Face2Face
- DeepFaceLab
- DeepfakeCapsuleGAN
- NeuralTextures

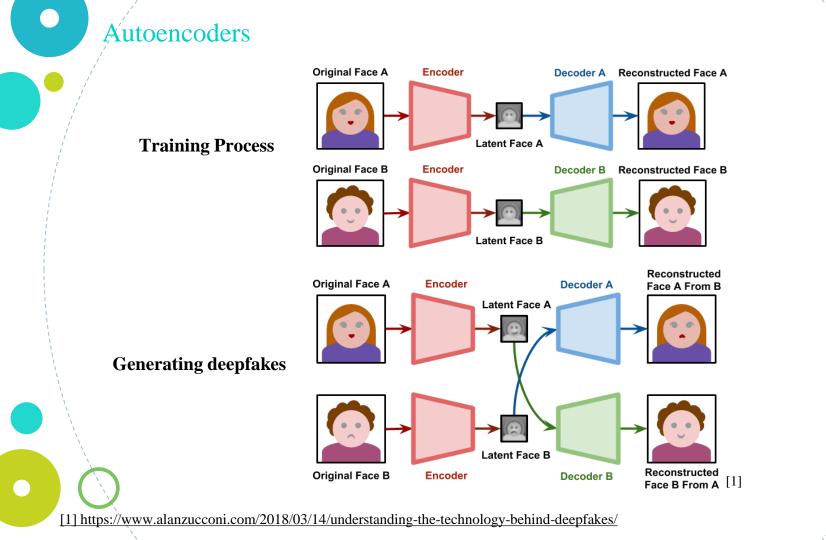
Các loại deepfake

- Face-swapping
- · Face re-enactment
- · Audio Deepfakes
- · Lip-syncing Deepfakes
- · Puppet-master

Cách thức tạo ra deepfake?

Autoencoders

- Nó bao gồm 2 phần: encoder và decoder
- Encoder được huấn luyện với nhiều bức hình của hai đối tượng mà người dùng muốn hoán đổi và trích xuất các đặc trưng nổi bật của các tấm hình.
- Decoder thì tạo lại và khôi phục hình ảnh từ những đặc điểm nổi bật của các tấm hình được cung cấp.
- Sau khi huấn luyện thì Encoder và Decoder sẽ khôi phục lại hai hình ảnh khác nhau



Cách thức tạo ra deepfake?

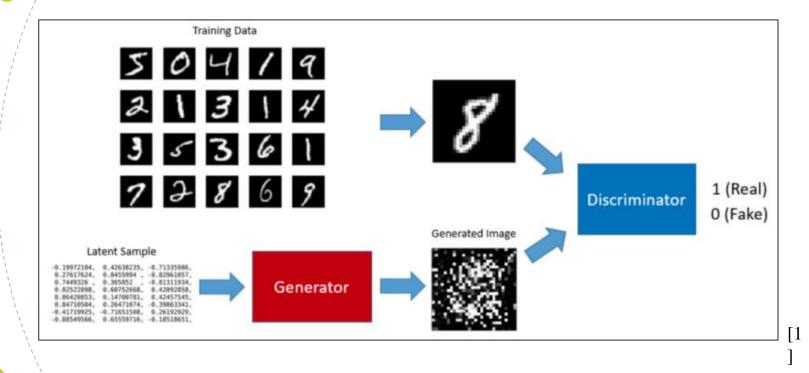
GANs

GANs bao gồm 2 neural networks và cùng "cạnh tranh" lẫn nhau:

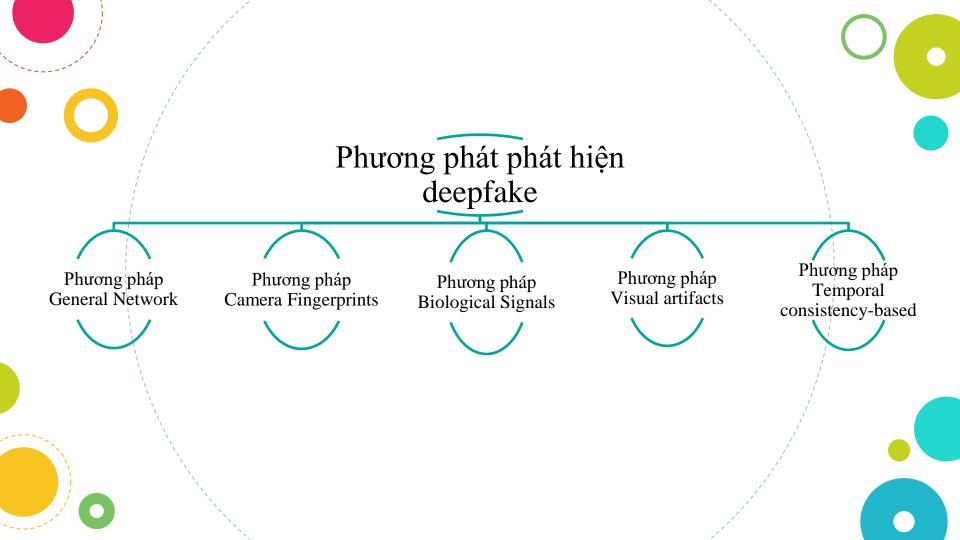
- Generator
- Discriminator

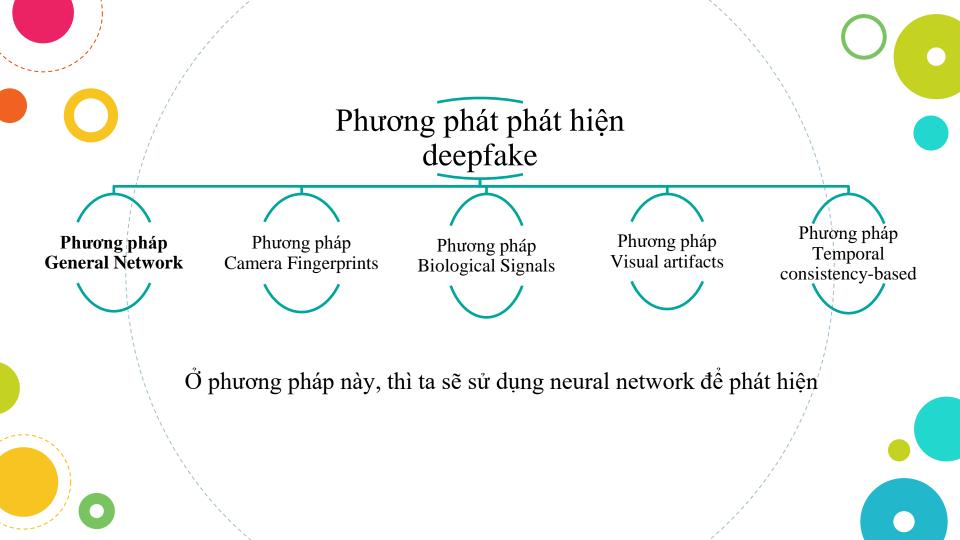
Generator cố gắng tạo ra dữ liệu giống thật nhất có thể, còn Discriminator cố gắng phân biệt dữ liệu mà Generator đổ vào và chứng minh nó là hàng fake sau đó thông báo lại cho Generator để nó cải thiện, cứ như thế quá trình này lặp đi lặp lại để generator có thể tạo ra sample hoàn hảo nhất mà Discriminator không thể phân biệt được.

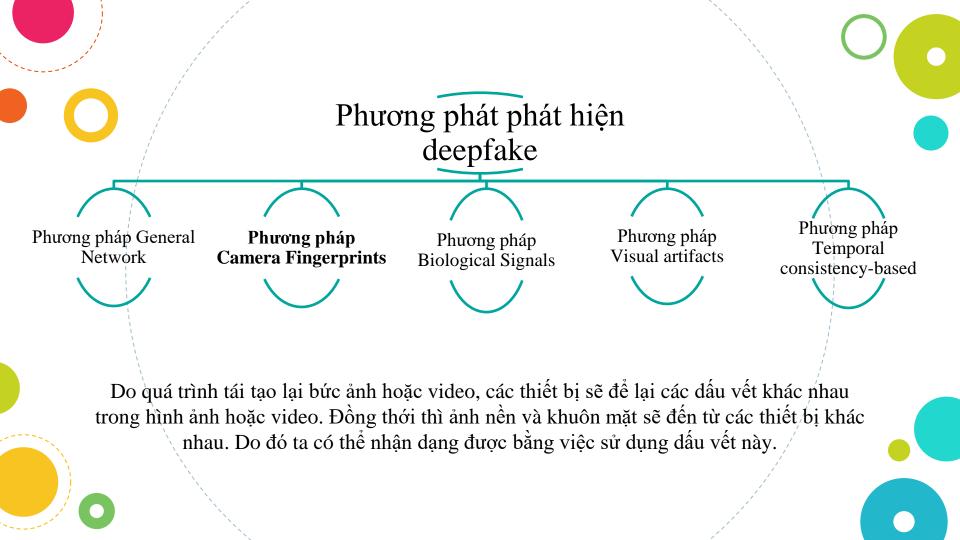
Generative Adversarial Networks

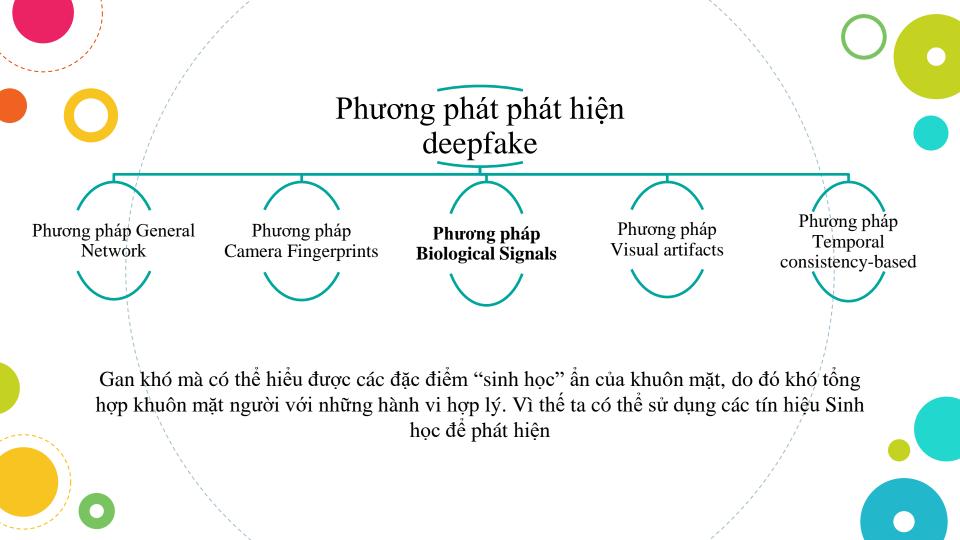


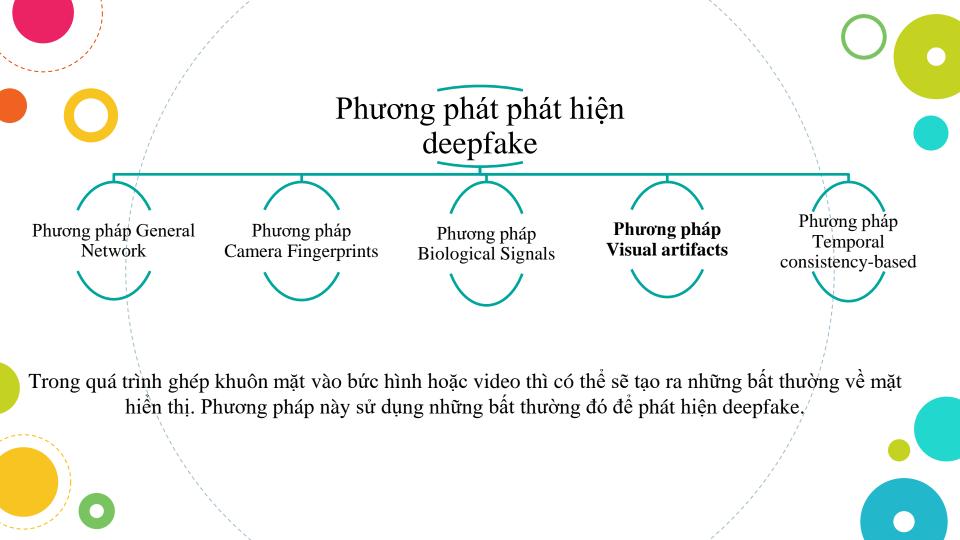


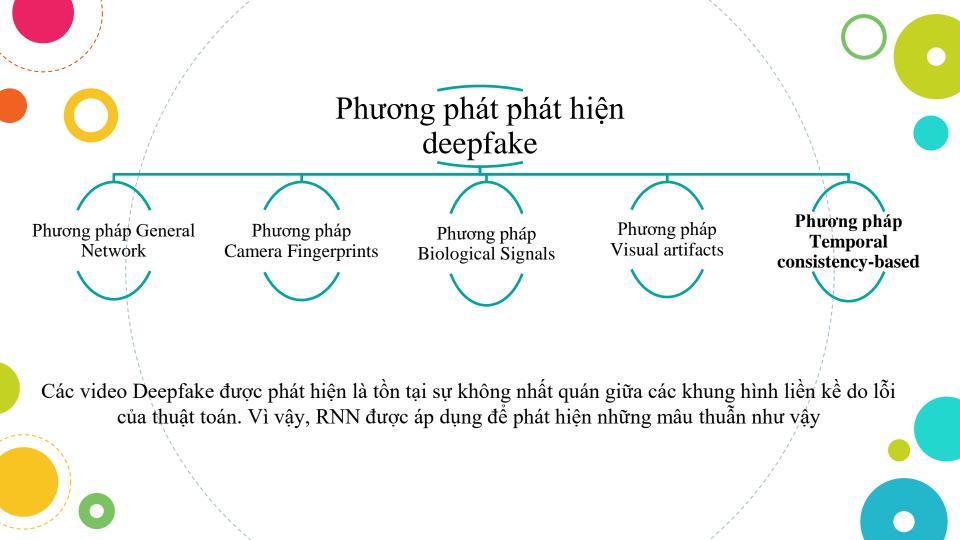


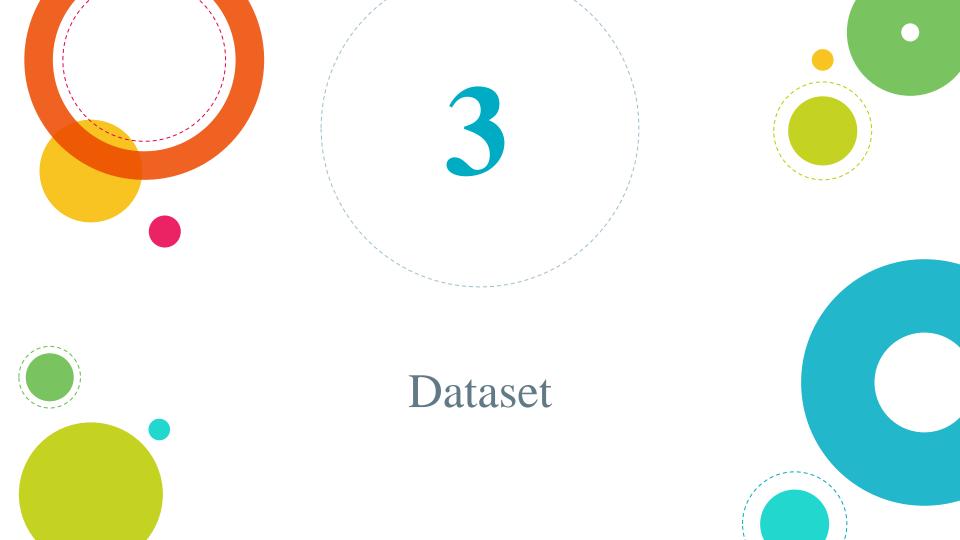


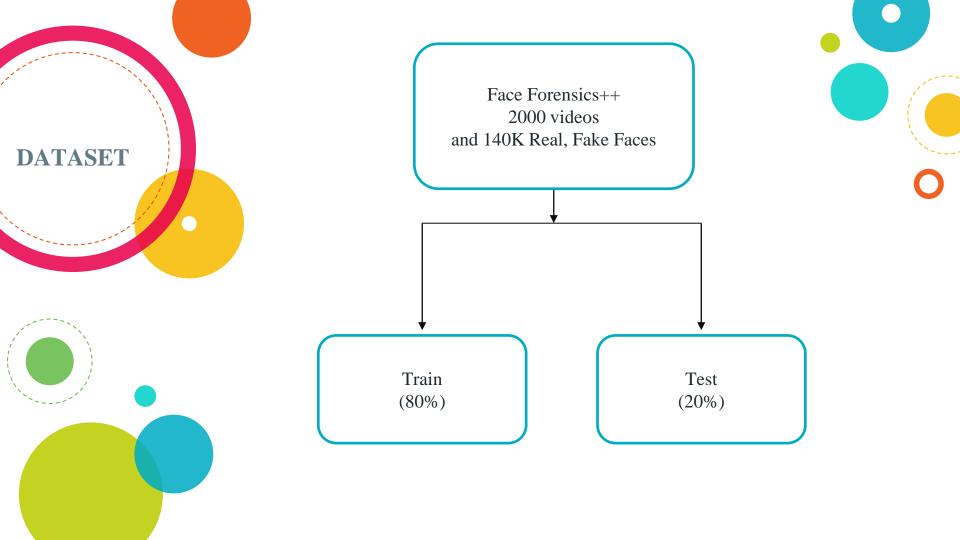




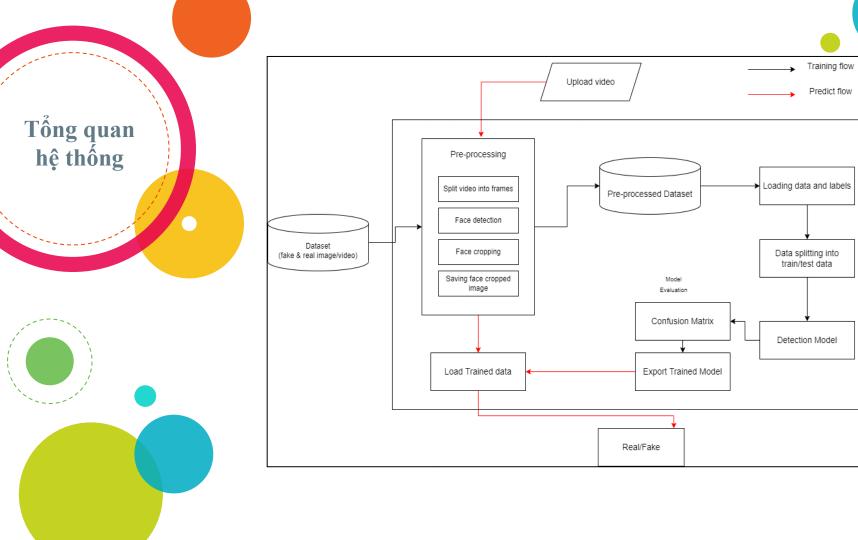


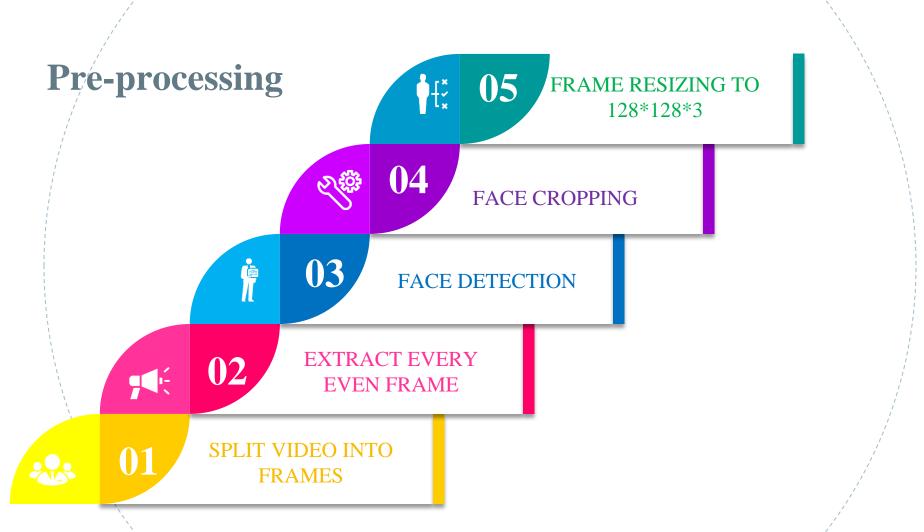


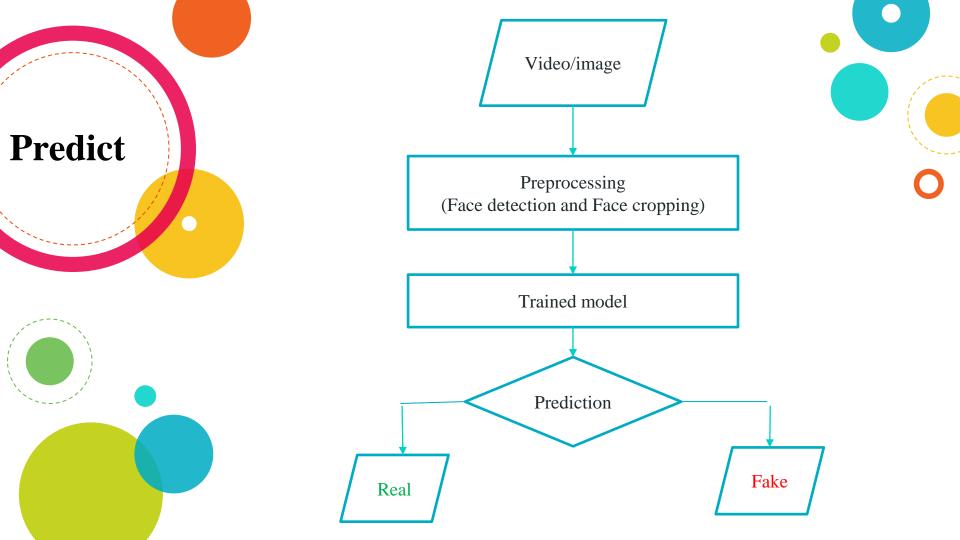




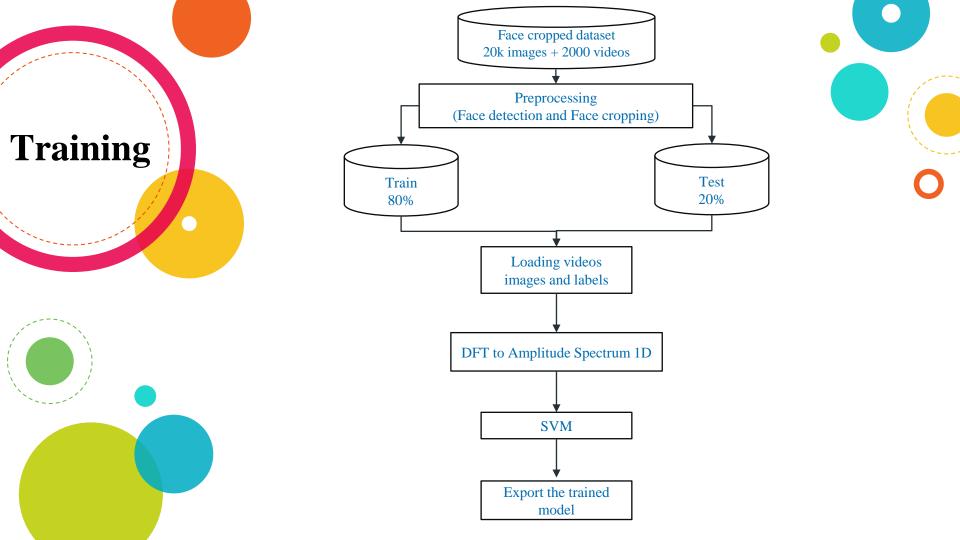


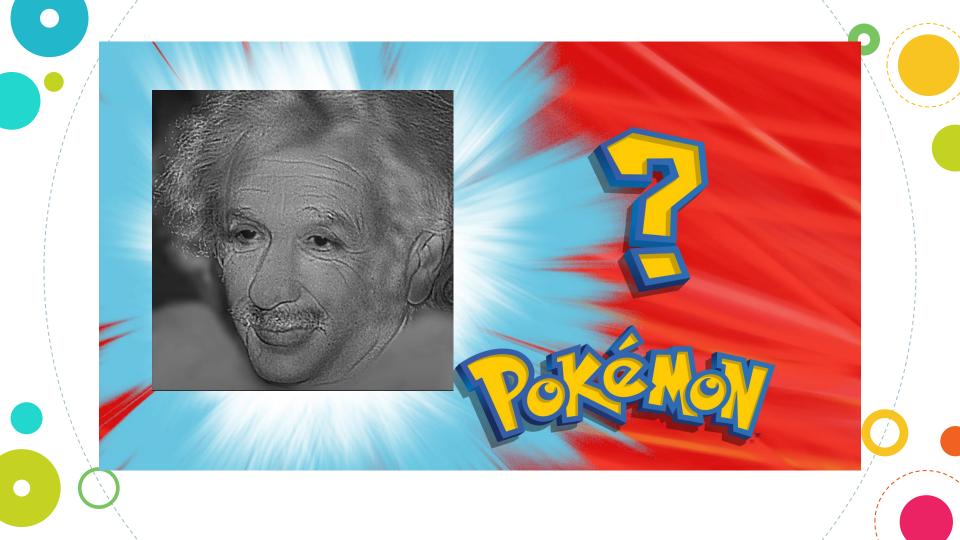




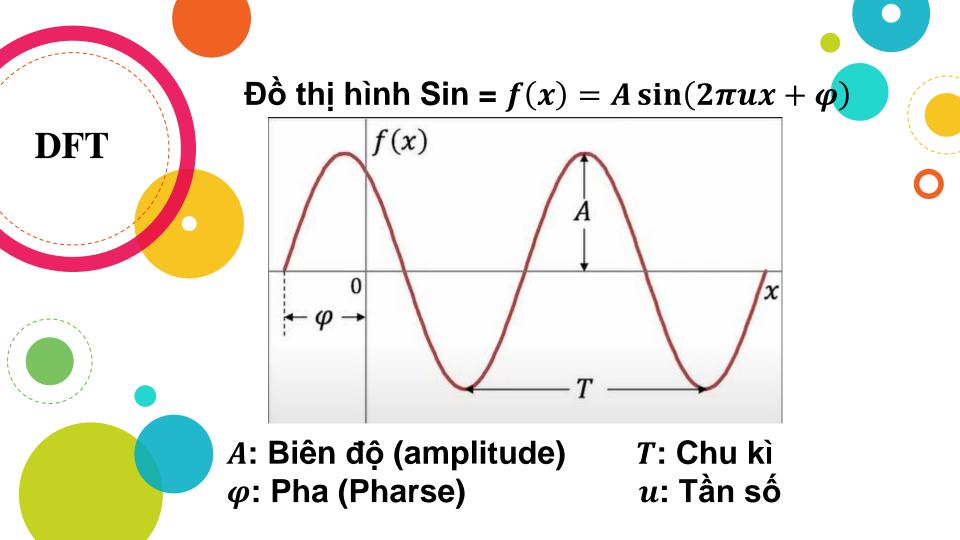


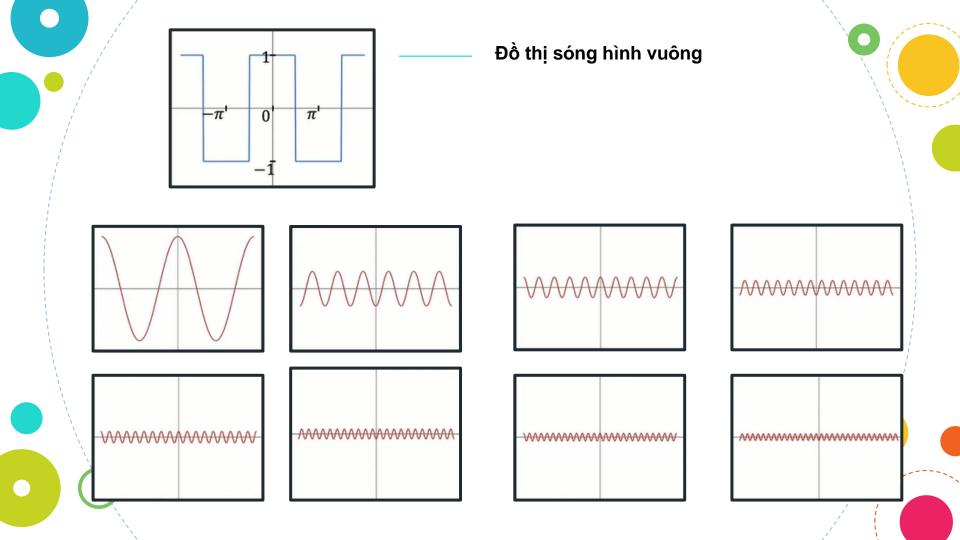


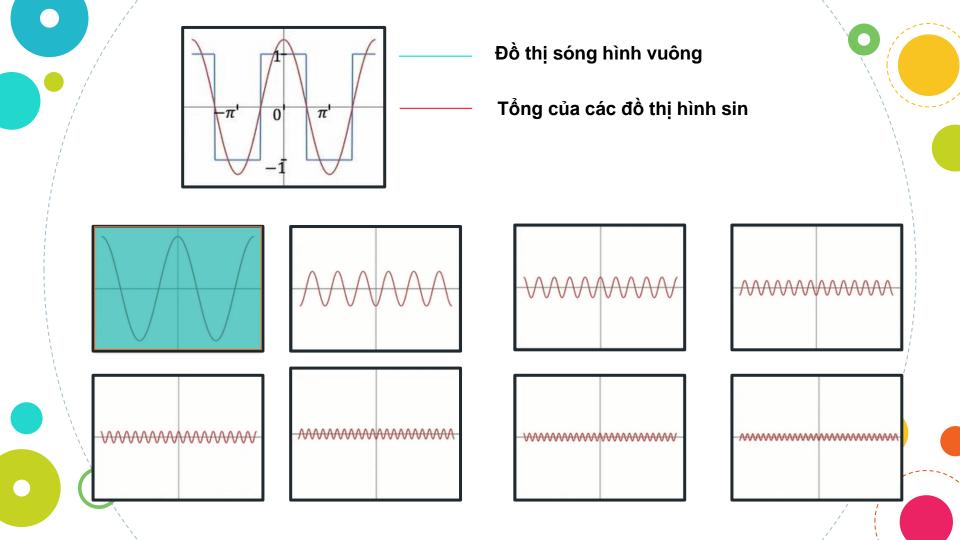


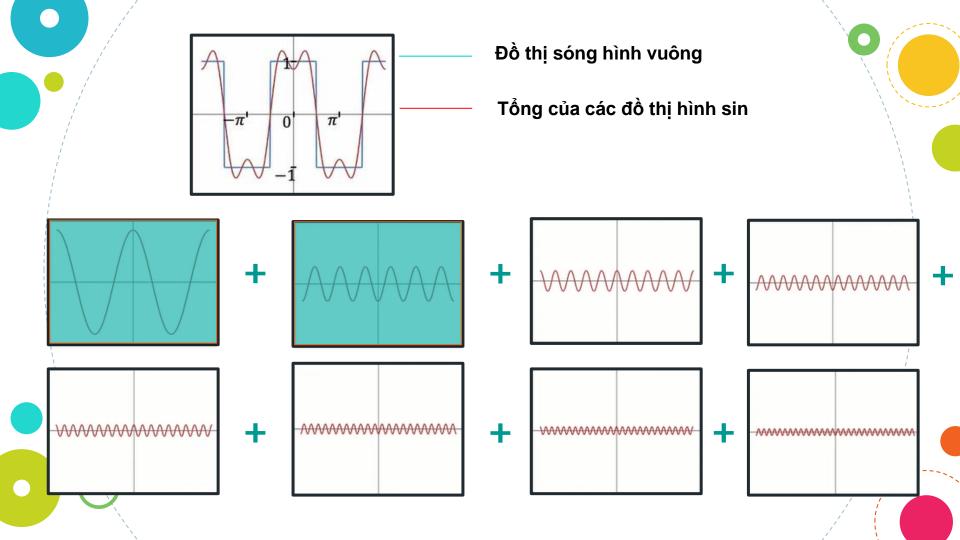


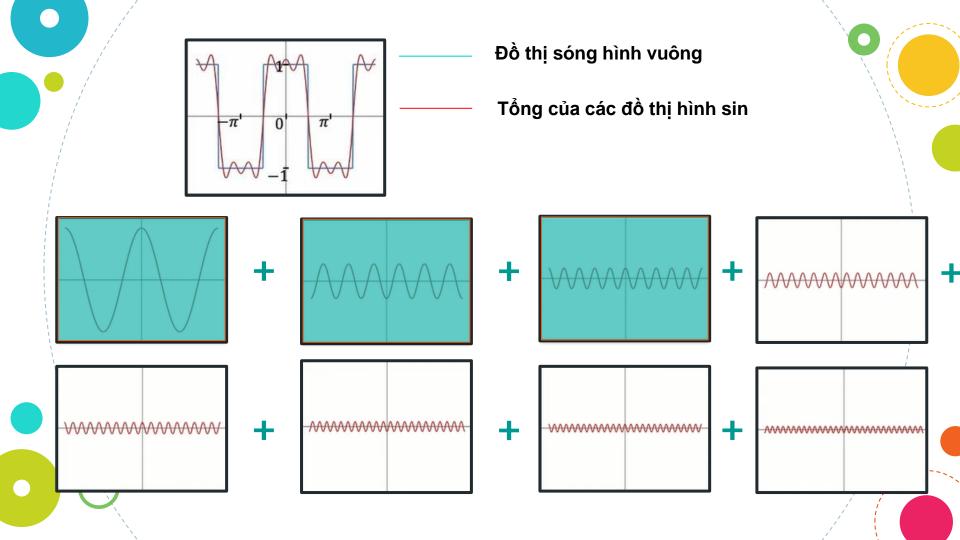


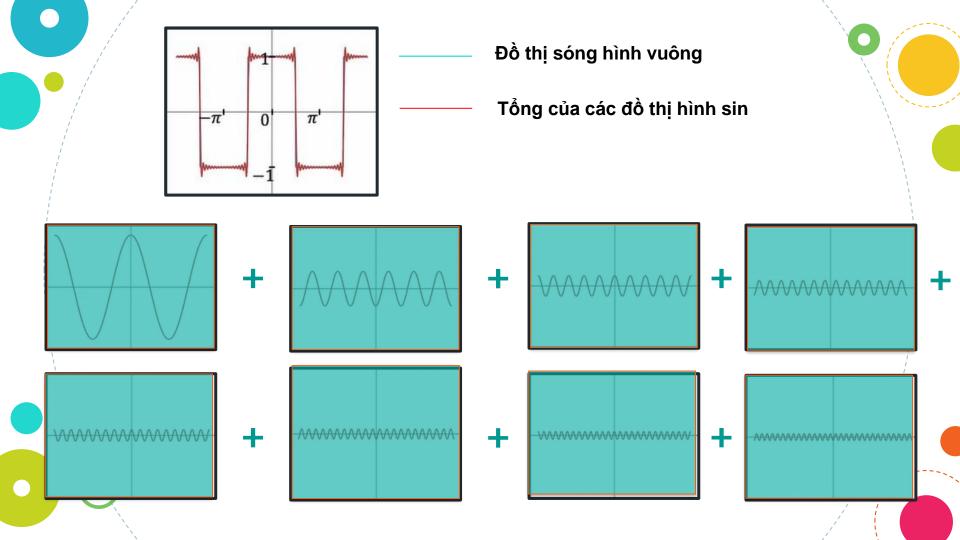


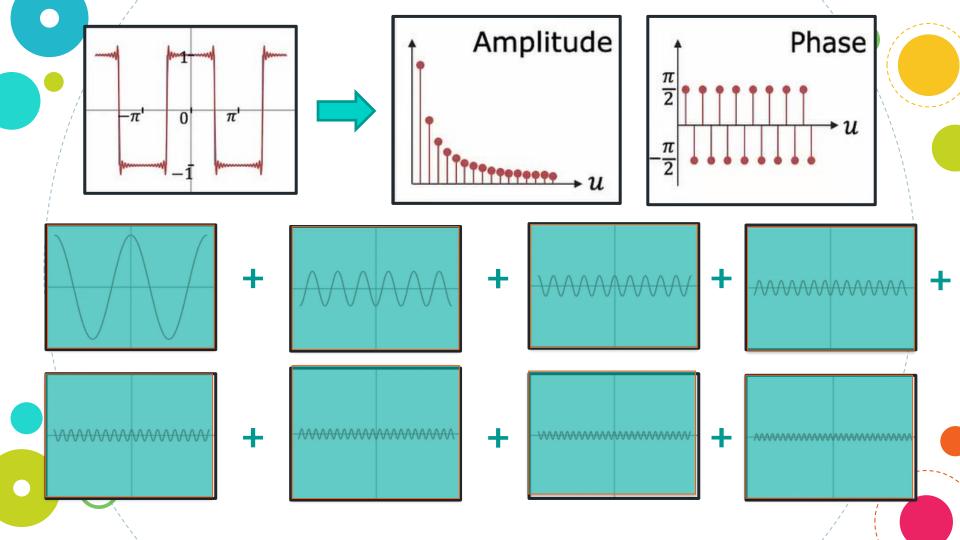


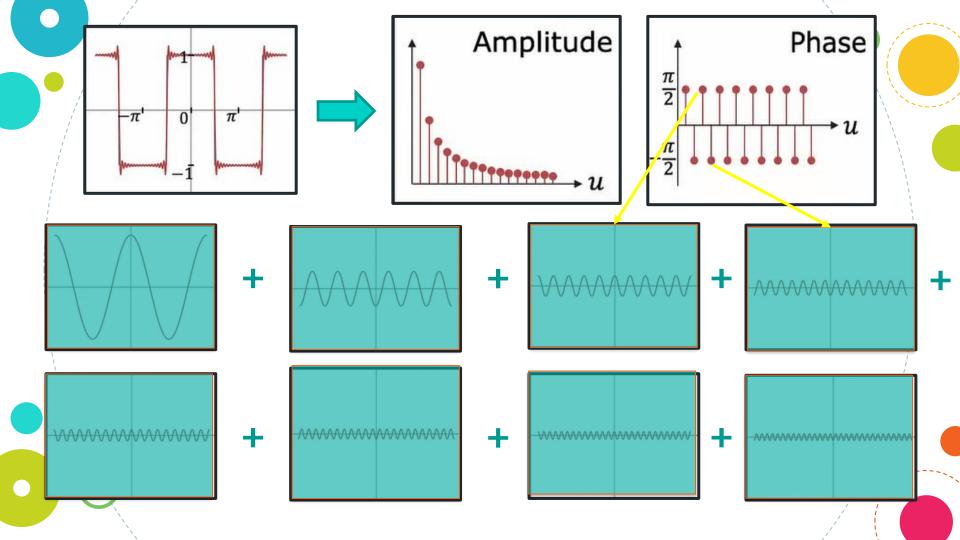


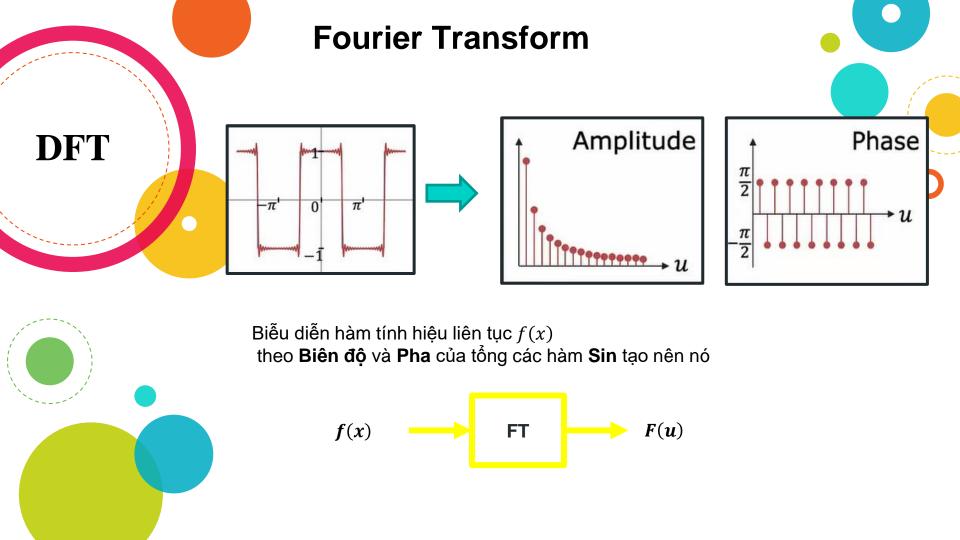


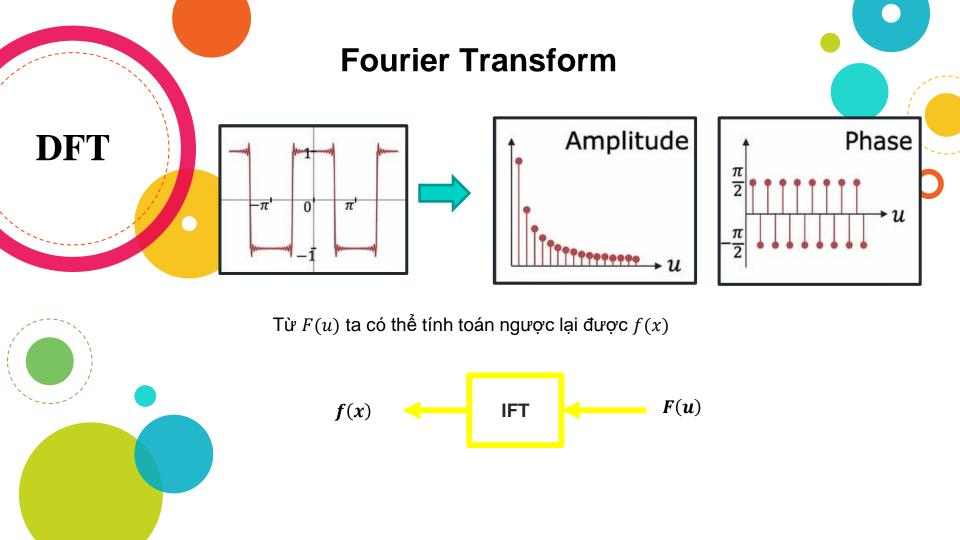












Fourier Transform:

$$F(u) = \int_{-\infty}^{\infty} f(x) e^{-i2\pi ux} dx$$

x : giá trị trong miền không gian

u: giá trị trong miền tần số

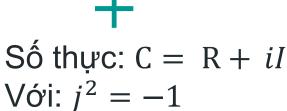
$$i = \sqrt{-1}$$

$$f(x) = \int_{-\infty}^{\infty} F(u) e^{i2\pi ux} du$$

Biểu diễn Fourier Transform sử dụng số phức

Công thức Euler:
$$e^{i\theta} = \cos \theta + i \sin \theta$$

F(u) = R(u) + i I(u)



Biên độ:
$$A(u) = \sqrt{R^2(u) + I^2(u)}$$

Pha:
$$\varphi(u) = \tan^{-1}\left(\frac{I(u)}{R(u)}\right)$$

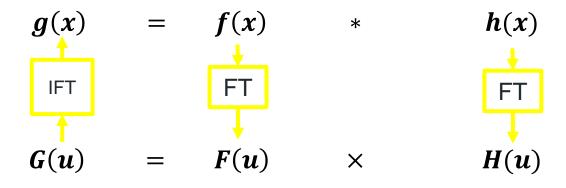
Công suất:
$$P(u) = R^2(u) + I^2(u)$$

Tính chất của Fourier Transform

Tính chất	Miền không gian	Miền tần số
Tuyến tính	$\alpha f_1(x) + \beta f_2(x)$	$\alpha F_1(u) + \beta F_2(u)$
 Tỉ lệ	f(ax)	$\frac{1}{ a } F(\frac{u}{a})$
 Phép dời	f(a-x)	$e^{-i2\pi u}F(u)$
Tính vi phân	$\frac{d^n}{dx^n}(f(x))$	$(i2\pi u)^n F(u)$

Convolution và Fourier Transform

Miền không gian	Miền tần số
g(x) = f(x) * h(x)	G(u) = F(u) H(u)
Phép tính chập	Phép nhân
g(x) = f(x) h(x)	G(u) = F(u) * H(u)
Phép nhân	Phép tính chập



2D Fourier Transform

$$F(u,v) = \iint_{-\infty}^{\infty} f(x) e^{-i2\pi(ux+vy)} dx$$

 $oldsymbol{u},oldsymbol{y}$ tương ứng là giá trị tần số dọc theo $oldsymbol{u},oldsymbol{v}$

2D Inverse Fourier Transform

$$f(x,y) = \iint_{-\infty}^{\infty} F(u,v)e^{i2\pi(ux+vy)}du$$

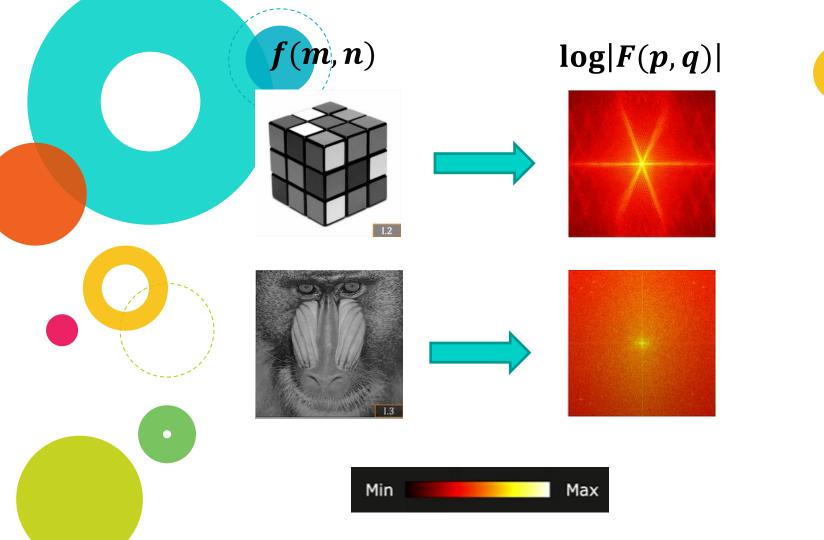
2D Discrete Fourier Transform:

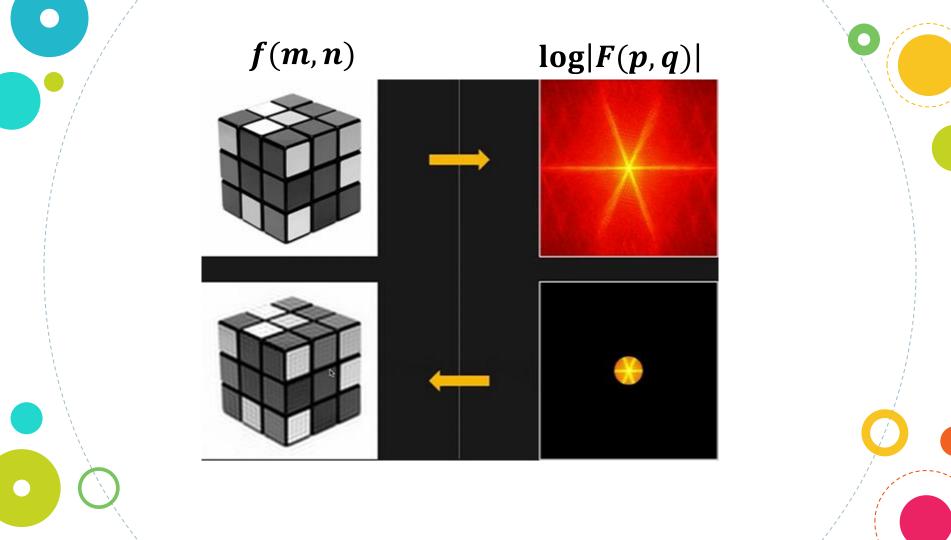
$$F[p,q] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f[m,n] e^{-i2\pi pm/M} e^{-i2\pi qn/N} \qquad p = 0 \dots M-1 q = 0 \dots N-1$$

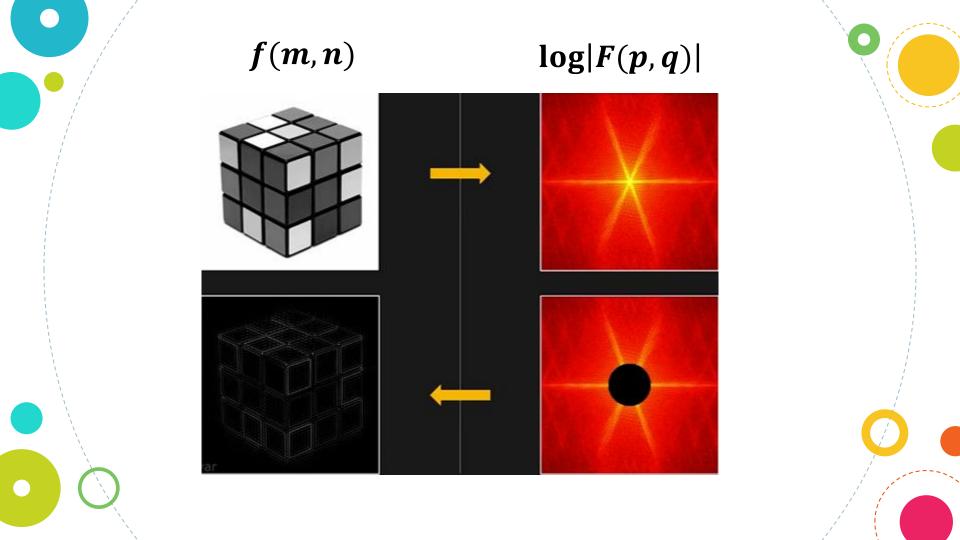
p, q tương ứng là giá trị tần số dọc theo M, N với ảnh đầu vào có kích thước $M \times N$

2D Inverse Dicrete Fourier Transform:

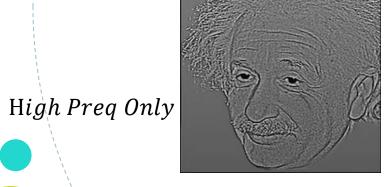
$$f[m,n] = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} F[p,q] e^{i2\pi pm/M} e^{i2\pi qn/N} \qquad m=0 \dots M-1 \\ n=0 \dots N-1$$

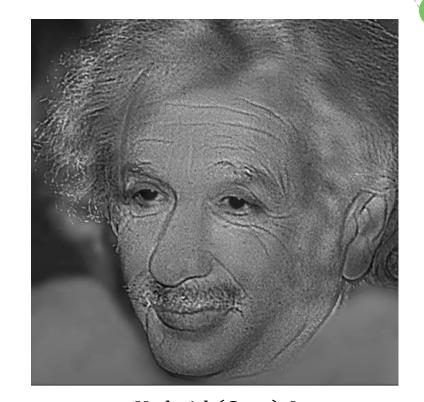










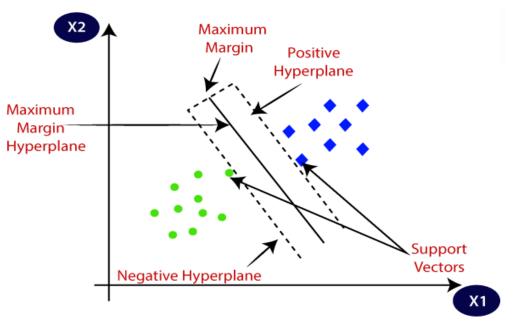


Hybrid (Sum) Image



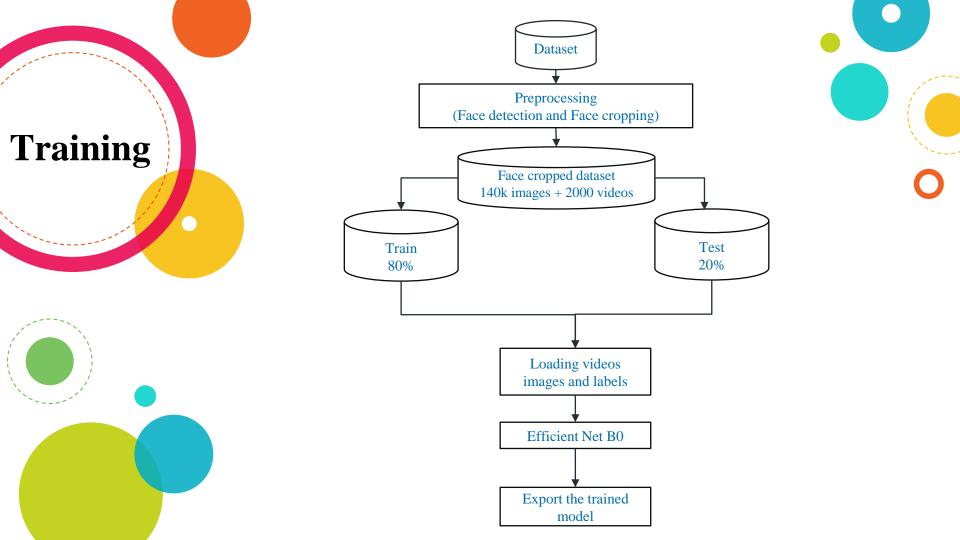
First Principles of Computer Vision https://www.youtube.com/watch?v=OOu5KP3Gvx0

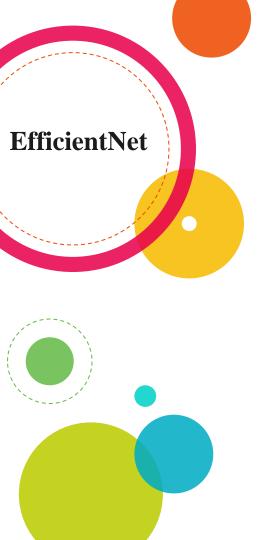




- Thuật toán học có giám sát sử dụng bài toán phân lớp(Phù hợp với dataset)
- Cho kết quả cao cho bài toán phân lớp dữ liệu có 2 lớp (Real và Fake)







- Mạng Nơ-ron tích chập (Convolutional Neural Networks - ConvNets) thường được phát triển với ngân sách tài nguyên cố định và sau đó được thu phóng để có độ chính xác tốt hơn nếu có nhiều tài nguyên hơn.

- Bởi vậy nên nhóm tác giả Mingxing Tan và Quoc V. Le đã nghiên cứu một cách có hệ thống và nhận thấy rằng việc cân bằng một cách có hệ thống độ sâu, chiều rộng và độ phân giải mạng có thể mang đến hiệu suất tốt hơn.

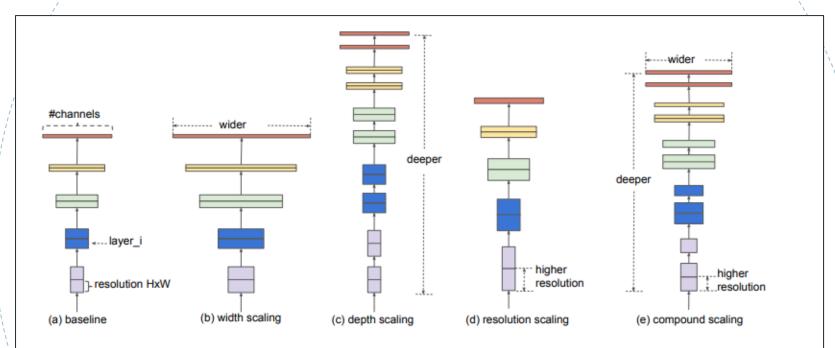
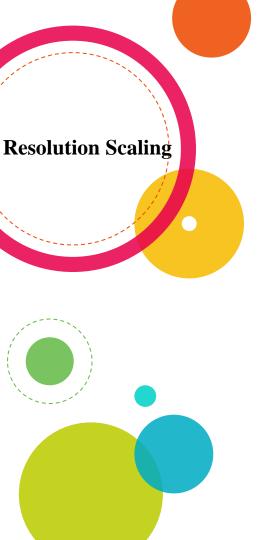


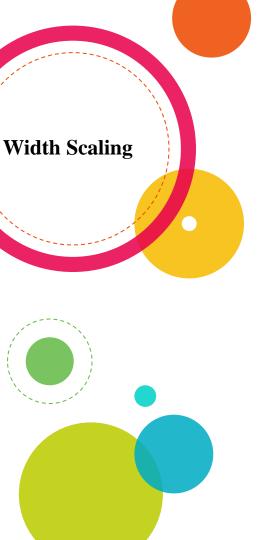
Figure 2. **Model Scaling.** (a) is a baseline network example; (b)-(d) are conventional scaling that only increases one dimension of network width, depth, or resolution. (e) is our proposed compound scaling method that uniformly scales all three dimensions with a fixed ratio.



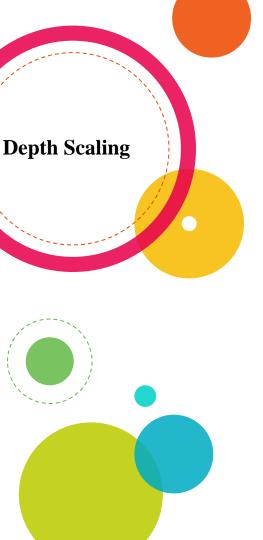
- Theo một cách trực quan, chúng ta có thể nói rằng trong một hình ảnh có độ phân giải cao, các đặc trưng sẽ có độ chi tiết cao hơn và do đó hình ảnh có độ phân giải cao sẽ hoạt động tốt hơn. Độ phân giải đầu vào cao hơn cung cấp hình ảnh chi tiết hơn và do đó nâng cao khả năng suy luận của mô hình về các đối tượng nhỏ hơn và trích xuất các mẫu mịn hơn







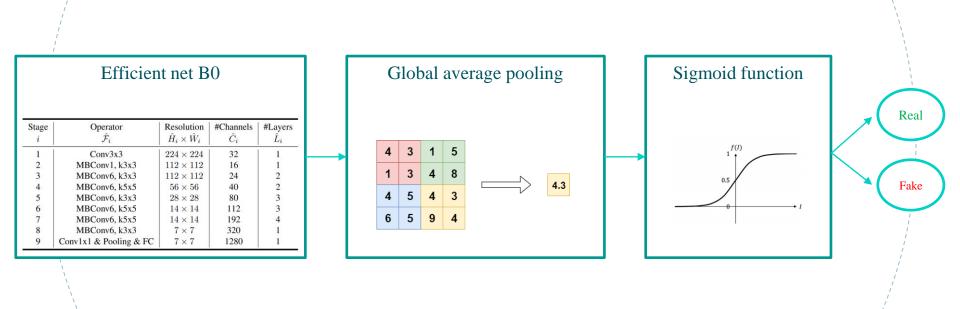
Việc thu phóng theo chiều rộng của mạng (theo như trong hình minh họa ta có thể hiểu là thêm dữ liệu đầu vào) cho phép các lớp tìm hiểu các tính năng chi tiết hơn. Tuy nhiên, cũng như trường hợp tăng chiều sâu, tăng chiều rộng ngăn cản mạng học các tính năng phức tạp, dẫn đến giảm độ chính xác.

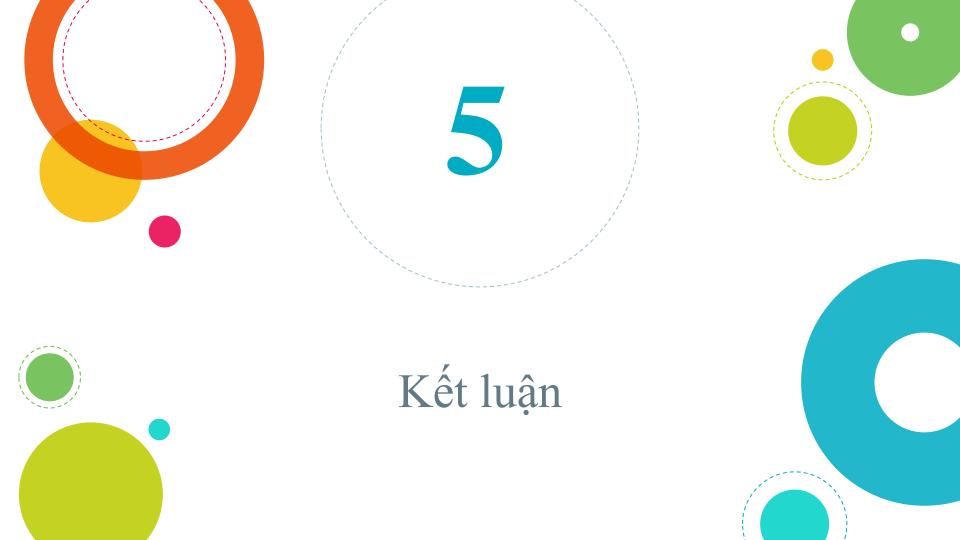


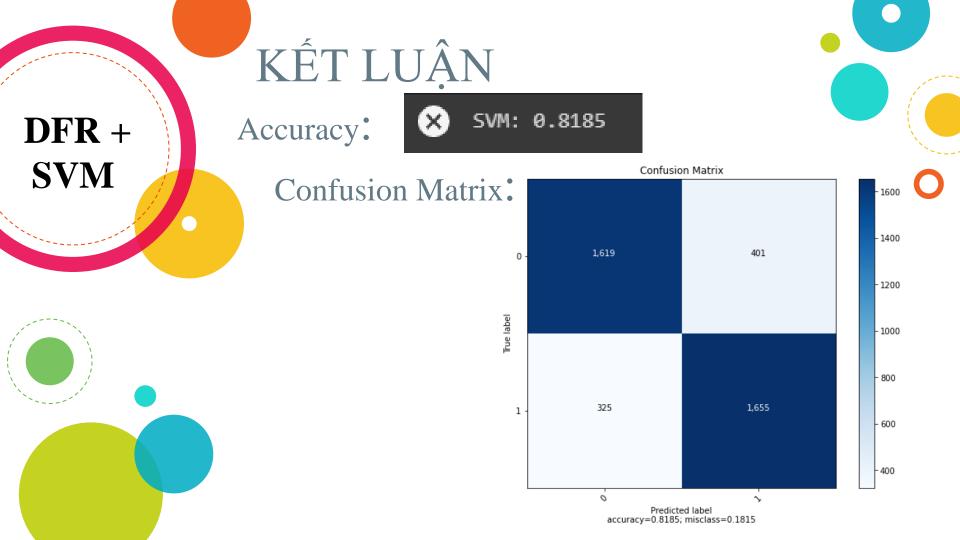
Tuy nhiên việc lạm dụng thu phóng theo chiều sâu có thể không cải thiện hiệu quả của mô hình, thậm chí có thể làm mô hình kém hiệu quả hơn so với mô hình ban đầu. Đúng là có một số lý do mà việc thêm nhiều lớp ẩn hơn sẽ cung cấp mức độ chính xác hơn cho mô hình. Tuy nhiên, điều này chỉ đúng với các tập dữ liệu lớn hơn, vì càng nhiều lớp với hệ số bước ngắn hơn sẽ trích xuất nhiều tính năng hơn cho dữ liệu đầu vào của bạn. Việc sử dụng một mô hình quá phức tạp với lượng dữ liệu không tương xứng, như ta đã biết, có thể gây ra hiện tượng Overfitting. Thêm nữa, các mạng sâu hơn có xu hướng bị **vanishing gradients** và trở nên khó đào tạo. Vậy nên không phải lúc nào, thu phóng theo chiều sâu cũng là sự lựa chọn thích hợp để cải thiện mô hình CNN.

- Vanishing gradients là vấn đề xảy ra khi huấn luyện các mạng nơ ron nhiều lớp. Khi huấn luyện, giá trị đạo hàm là thông tin phản hồi của quá trình lan truyền ngược. Các cập nhật thực hiện bởi Gradient Descent không làm thay đổi nhiều weights của các layer đó, khiến chúng không thể hội tụ và DNN sẽ không thu được kết quả tốt

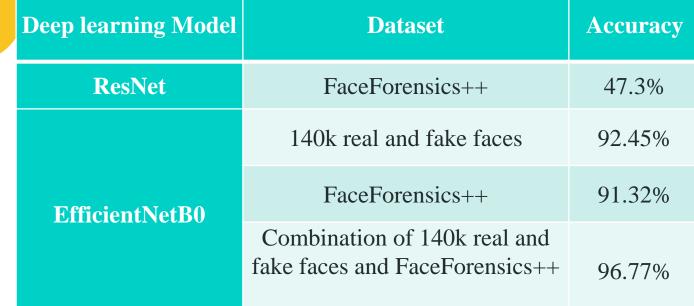
Kiến trúc model











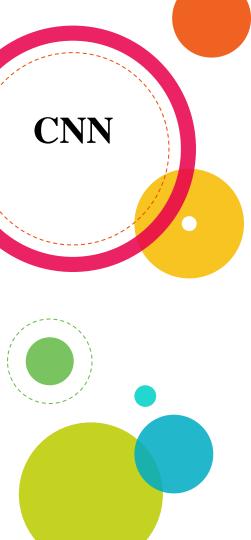




THỰC NGHIỆM HƯỚNG PHÁT TRIỂN

- -Acc cao nghiêng về face có những vết nhòe bất thường
- Chưa detect được Deepfake dùng công nghệ,
 kỹ thuật cao
- Mở rộng Dataset (mặt người từ nhiều quốc gia, dùng Deepfake công nghệ cao để tạo thêm fake data)





THỰC NGHIỆM HƯỚNG PHÁT TRIỂN

- Model dự đoán khá chính xác với những người châu Âu và châu Mỹ.
- Tuy nhiên đối với những video chứa dân Châu Á thì model dự đoán ko được chính xác lắm tại dataset hơi ít mặt người châu Á.
- Model dự đoán với những video sử dụng công nghệ kỹ thuật deepfake tiên tiến thì predict tương đối đúng.
- Mở rộng Dataset (mặt người từ nhiều quốc gia, dùng Deepfake công nghệ cao để tạo thêm fake data)
- Nâng cấp là thêm phần chọn khuôn mặt trong video mà mình detect là real hay fake và sẽ thêm tính năng nhận diện real time.

