# BẢO VỆ NGƯỜI NỔI TIẾNG KHỎI DEEPFAKE VỚI IDENTITY CONSISTENCY TRANSFORMER

**Nguyễn Quốc Khánh**[1,4] **Nguyễn Trần Minh Anh**[2,4] **Lê Nguyễn Bảo Hân**[3,4]

{[1]20521452, [2]20520394, [3]20520174}@gm.uit.edu.vn [4]Trường Đại học Công nghệ Thông tin ĐHQG TP. HCM

## What

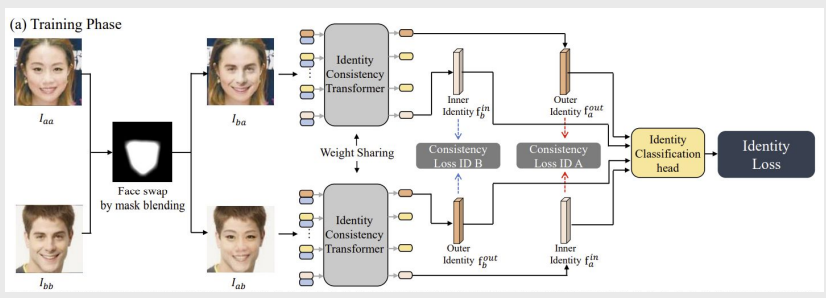
- **Identity Consistency Transformer** (ICT) là một phương pháp phát hiện khuôn mặt giả mạo dựa trên **thông tin ngữ nghĩa cấp cao**, đặc biệt là thông tin nhận dạng
- Phát hiện khuôn mặt khả nghi bằng cách tìm sự không nhất quán ở các vùng trong và ngoài mặt

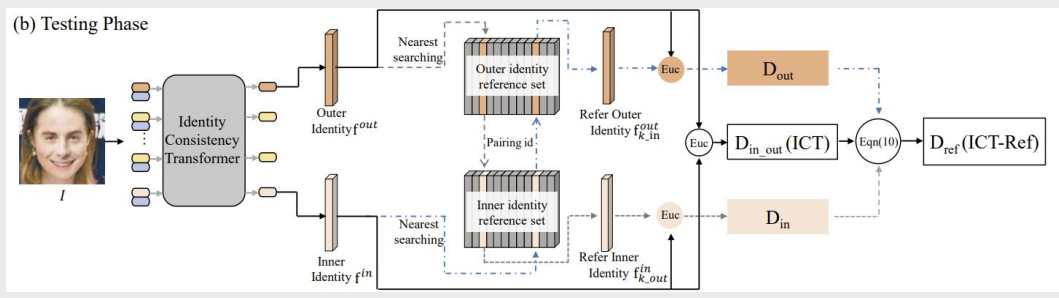
## Why

- Việc sử dụng và lan truyền nội dung Deepfake một cách độc hại đã làm dấy lên nỗi lo sợ nghiêm trọng trong xã hội
- Hầu hết các phương pháp hiện có chỉ khai thác các **đặc trưng cấp thấp**
- Đầu vào là video frame thường bị giảm chất lượng hình ảnh, không ổn định

## Overview

Identity Extraction Model → Identity Consistency Detection




## Description

### 1. Identity Extraction Model

- Áp dụng mô hình **Vision Transformer** (ViT) cho bài toán phân loại khuôn mặt
- Bổ sung ***Consistency Loss*** để phù hợp hơn với bài toán phát hiện nhất quán danh tính
- Bộ mã hóa gồm nhiều block xếp chồng lên nhau. Mỗi block bao gồm một lớp **Multi-Head Self-Attention** và một lớp **MLP** chuẩn hóa ở phía trước

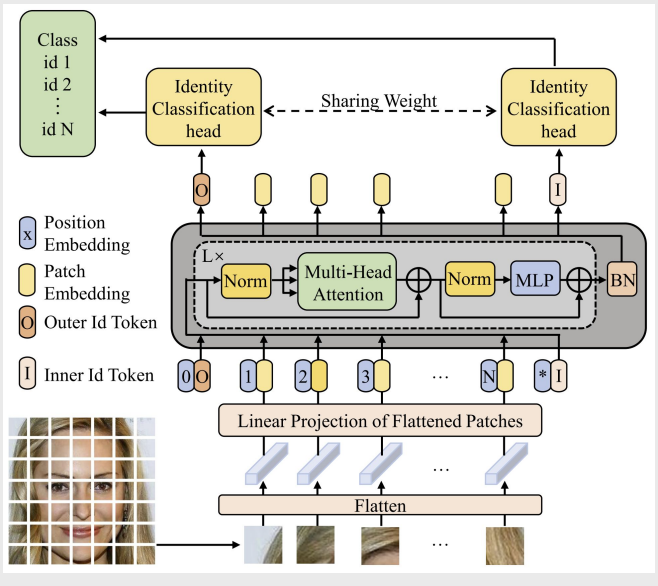



***Figure 1***. Architecture of the identity extraction model

### 2. Identity Consistency Detection

- Sử dụng độ đo khoảng cách giữa 2 vector inner và outer được trích xuất. $D_{in\_out}$ nhỏ cho thấy khả năng danh tính càng nhất quán

$$D_{in\_out} = d(\mathbf{f}^{in}, \mathbf{f}^{out}).$$

- Xây dựng bộ tham chiếu chứa các cặp vector nhận dạng của tất cả hình ảnh thực có sẵn. Sau đó, sử dụng inner identity để tìm **nearest neighbor** trong tập tham chiếu. Lấy outer identity tương ứng trong tập tham chiếu và tính khoảng cách với outer identity của hình ảnh đáng ngờ

$$D_{out} = d(\mathbf{f}^{out}, \mathbf{f}^{out}_{k^{in}}). \quad \alpha_{in} = d(\mathbf{f}^{in}, \mathbf{f}^{in}_{k^{in}})$$

- Tương tự với inner identity, ta có:

$$D_{in} = d(\mathbf{f}^{in}, \mathbf{f}^{in}_{k^{out}}), \quad \alpha_{out} = d(\mathbf{f}^{out}, \mathbf{f}^{out}_{k^{out}}).$$

- Reference-assisted identity consistency detection: kết hợp mọi khoảng cách

$$D_{ref} = \lambda D_{in\_out} + \omega(\alpha_{in}) D_{out} + \omega(\alpha_{out}) D_{in},$$

### 3. Benefits, Limitations and Impacts

***Benefits:***

- Tính tổng quát hóa cho mọi sự biến đổi khuôn mặt khác nhau
- Ít bị ảnh hưởng bởi sự suy giảm chất lượng hình ảnh
- Sử dụng kỹ thuật nhận dạng khuôn mặt tiên tiến, thông tin trích xuất đáng tin cậy

***Limitations***

- Phương pháp chủ yếu chú trọng face swapping và có thể thất bại trong phát hiện kết quả tái hiện khuôn mặt mà danh tính được dự định giữ nguyên

***Impacts***

- Chỉ cần một nhãn one-hot cho mỗi người và không cần tên. Phù hợp với mục tiêu bảo vệ thông tin nhận dạng khỏi việc sử dụng Deepfake với mục đích xấu