

**VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY  
UNIVERSITY OF INFORMATION TECHNOLOGY**



**BÁO CÁO ĐỒ ÁN CUỐI KỲ**

**MÔN HỌC: CS519 - PHƯƠNG PHÁP LUẬN NGHIÊN CỨU KHOA HỌC**

**GIẢNG VIÊN: PGS.TS. LÊ ĐÌNH DUY**

**THỜI GIAN: 09/2022 - 12/2022**

**THÀNH VIÊN NHÓM:**

**NGUYỄN QUỐC KHÁNH - 20521452**


**NGUYỄN TRẦN MINH ANH - 20520394**

**LÊ NGUYỄN BẢO HÂN - 20520174**

**Hồ Chí Minh, 25 tháng 02 năm 2023**

# THÔNG TIN CHUNG CỦA NHÓM AKH

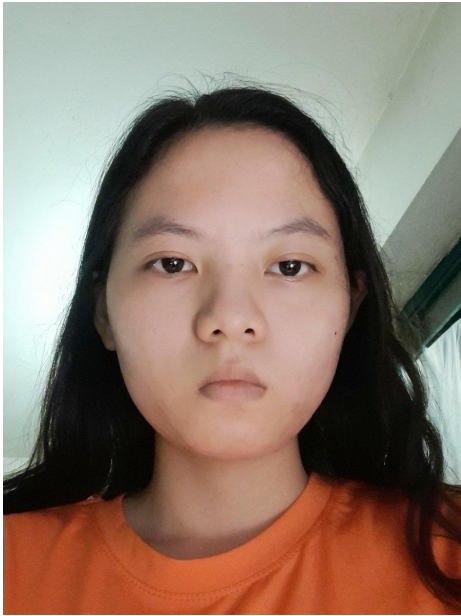
- Link YouTube video của báo cáo (tối đa 5 phút):  
<https://www.youtube.com/watch?v=LxhwHdXHndw>
- Link slides (dạng .pdf đặt trên Github của nhóm):  
(ví dụ: <https://github.com/mynameuit/CS519.M1.KHCL/TenDeTai.pdf>)

<ul style="list-style-type: none"><li>● Họ và Tên: Nguyễn Quốc Khánh</li><li>● MSSV: 20521452</li></ul> 	<ul style="list-style-type: none"><li>● Lớp: CS519.N11</li><li>● Tự đánh giá (điểm tổng kết môn): 9/10</li><li>● Số buổi vắng: 1</li><li>● Số câu hỏi QT cá nhân: 3</li><li>● Số câu hỏi QT của cả nhóm: 5</li><li>● Link Github: <a href="https://github.com/nqkhanh2002/CS519.N11">https://github.com/nqkhanh2002/CS519.N11</a></li><li>● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none"><li>○ Chọn ý tưởng đồ án</li><li>○ Viết báo cáo đồ án</li><li>○ Quay video báo cáo, chỉnh sửa poster, slide</li></ul></li></ul>
--	--

<ul style="list-style-type: none"><li>● Họ và Tên: Nguyễn Trần</li></ul>	<ul style="list-style-type: none"><li>● Lớp: CS519.N11</li></ul>
--	--

Anh Minh

- MSSV: 20520394



- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 3
- Số câu hỏi QT của cả nhóm: 5
- Link Github:  
<https://github.com/nqkhanh2002/CS519.N11>
- Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
  - Lên ý tưởng đồ án
  - Viết báo cáo đồ án, chỉnh sửa slide và poster
  - Làm video YouTube

- Họ và Tên: Lê Nguyễn Bảo Hân

- MSSV: 20520174



- Lớp: CS519.N11
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 3
- Số câu hỏi QT của cả nhóm: 5
- Link Github:  
<https://github.com/nqkhanh2002/CS519.N11>
- Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
  - Lên ý tưởng đồ án
  - Viết báo cáo đồ án, chỉnh sửa slide và poster, video youtube

	○ Chính sửa hoàn thiện
--	------------------------

# ĐỀ CƯƠNG NGHIÊN CỨU

**TÊN ĐỀ TÀI (IN HOA)**

BẢO VỆ NGƯỜI NỔI TIẾNG KHỎI DEEPPFAKE VỚI IDENTITY  
CONSISTENCY TRANSFORMER

**TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**

PROTECTING CELEBRITIES FROM DEEPPFAKE WITH IDENTITY  
CONSISTENCY TRANSFORMER

**TÓM TẮT** *(Tối đa 400 từ)*

Nghiên cứu này đề xuất một hướng tiếp cận để bảo vệ người nổi tiếng khỏi các cuộc tấn công Deepfake bằng cách sử dụng phương pháp Identity Consistency Transformer (ICT), một phương pháp phát hiện khuôn mặt giả mạo dựa trên thông tin ngữ nghĩa cấp cao, đặc biệt là thông tin nhận dạng và phát hiện khuôn mặt khả nghi bằng cách tìm sự không nhất quán ở các vùng trong và ngoài mặt. Phương pháp này kết hợp tính toán hàm mất mát nhất quán (*consistency loss*) thay vì chỉ áp dụng hàm mất mát cho nhận diện khuôn mặt.

Về dự định, nhóm báo cáo sẽ thử nghiệm trên ba bộ dữ liệu MS-Celeb-1M, FaceForensics++ và Celeb-DF để đánh giá khả năng khái quát hóa của mô hình và khả năng ứng dụng trong thực tế. Độ đo sử dụng chủ yếu là AUC (area under the Receiver Operating Characteristic curve) và báo cáo kết quả ở cấp khung hình cho tất cả thử nghiệm. Cuối cùng, nhóm báo cáo mong muốn xây dựng ứng

dụng nhằm trực quan hóa kết quả nghiên cứu cũng như áp dụng cho thực tiễn.

Về ý nghĩa, đề tài có ý nghĩa quan trọng trong việc bảo vệ quyền riêng tư và danh tiếng của những người nổi tiếng, vì nó có thể được tích hợp vào các hệ thống bảo mật để xác thực hình ảnh và video, ngăn chặn các tác nhân độc hại thao túng và phổ biến nội dung Deepfake. Nghiên cứu này góp phần vào những nỗ lực không ngừng nhằm chống lại mối đe dọa ngày càng tăng của công nghệ Deepfake và bảo vệ quyền riêng tư cũng như bảo mật của các cá nhân trong thời đại kỹ thuật số.

### **GIỚI THIỆU** *(Tối đa 1 trang A4)*

Kỹ thuật Deepfake [3,4] đã được phát triển rộng rãi để có thể tạo ra những bức ảnh hoặc video giả vô cùng chân thật với khuôn mặt được thay thế với ai đó trong bức ảnh khác. 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 và niềm tin của chúng ta với các phương tiện truyền thông trực tuyến. Do đó, phát hiện khuôn mặt giả mạo là một nhu cầu cấp thiết và đã thu hút được sự quan tâm đáng kể trong thời gian gần đây. Vì vậy, bài toán phát hiện khuôn mặt giả mạo - Deepfake Detection đã ra đời.

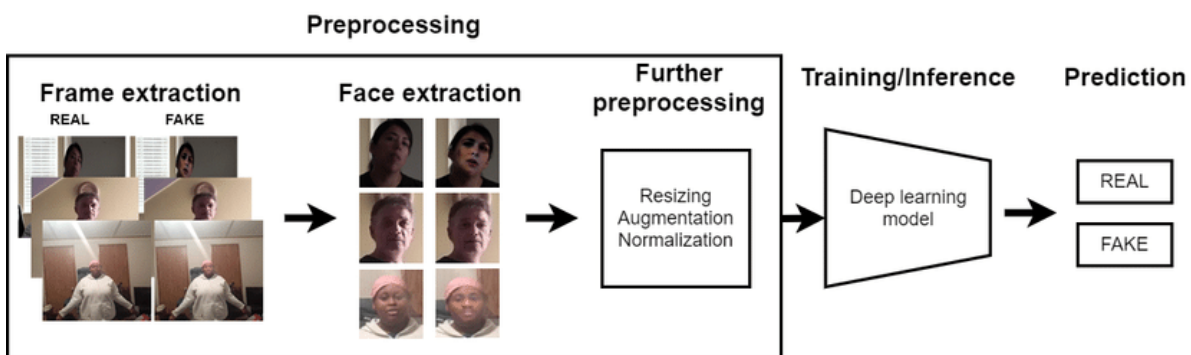
Gần đây, nhiều nỗ lực dành cho việc phát hiện khuôn mặt giả mạo đã đạt được hiệu suất đầy hứa hẹn. Hầu hết các phương pháp hiện có phân biệt hình ảnh giả bằng cách khai thác các đặc trưng cấp thấp và tìm kiếm các đối tượng dữ liệu ẩn được tạo ra. Tuy nhiên, các phương pháp này tồn tại hai vấn đề phổ biến: (1) phát hiện Deepfake thường được thực hiện trên các video nhưng các video frame thường bị giảm chất lượng hình ảnh (chẳng hạn như tỷ lệ thay đổi, nhiễu và chuyển đổi codec video,...); và (2) khi nội dung Deepfake được tạo trở nên chân thực một cách thuyết phục, dấu vết giả mạo ở ngữ nghĩa cấp thấp trở nên rất khó

phát hiện. Điều này làm cho khả năng phát hiện Deepfake không ổn định với đầu vào là video. Nghiên cứu này đề xuất phát hiện Deepfake bằng cách sử dụng hình ảnh đầu vào để nhận được nhiều thông tin nhận dạng quan trọng về mặt ngữ nghĩa.

Với bài toán này, dữ liệu đầu vào sẽ bao gồm bức ảnh cá nhân người dùng - source model (bức ảnh này sẽ bao gồm khuôn mặt). Mục tiêu đầu ra sẽ là phân loại ảnh đầu vào là hình ảnh thực hay hình ảnh được tạo bởi Deepfake.

**Input:** Hình ảnh khuôn mặt cần xác thực

**Output:** Phân loại hình ảnh là một hình ảnh thực hay hình ảnh được tạo bởi Deepfake



**Hình 1:** Ví dụ về một pipeline cho bài toán Deepfake Detection

Đáng chú ý, phần lớn đối tượng bị giả mạo khuôn mặt là các chính trị gia, người nổi tiếng hay lãnh đạo, vì ảnh và video của họ xuất hiện rất nhiều trên internet, dẫn đến dễ dàng bị thao túng để tạo ra nội dung Deepfake với độ chân thực ấn tượng. Tuy nhiên, các thuật toán phát hiện trước đây chỉ đưa ra dự đoán về sự giả mạo dựa trên các hình ảnh đáng ngờ và bỏ qua việc khai thác những dữ liệu có sẵn miễn phí đó. Vì vậy, nhóm chúng tôi đã đề xuất phương pháp phát hiện video Deepfake về người nổi tiếng bằng cách sử dụng mô hình dựa trên **Transformer: Identity Consistency Transformer**.

## **MỤC TIÊU** (*Viết trong vòng 3 mục tiêu*)

1. Khám phá, khảo sát các hướng tiếp cận gần đây dựa trên kỹ thuật học máy và thị giác máy tính cho bài toán phát hiện Deepfake. Xem xét và đánh giá độ hiệu quả của các phương pháp này trong việc phát hiện các trường hợp Deepfake của người nổi tiếng.
2. Xây dựng mô hình phát hiện giả mạo khuôn mặt có tên là Identity Consistency Transformer (ICT) dựa trên thông tin ngữ nghĩa cấp cao và đạt kết quả tốt hơn so với các mô hình hiện nay về độ chính xác và tốc độ xử lý, thực nghiệm và đánh giá trên các bộ dữ liệu tiêu chuẩn để đưa ra so sánh.
3. Nghiên cứu, thực nghiệm và đề xuất cải tiến thuật toán Transformer hiện có để kiểm chứng hiệu quả với module bài toán phân loại khuôn mặt. Cụ thể, thực nghiệm để xem xét liệu Transformer với cơ chế global attention có thể học được các đặc trưng ngữ nghĩa quan trọng cho việc phân loại. Cải tiến Transformer bằng việc thêm vào một hàm mất mát nhất quán (consistency loss) để phù hợp hơn với bài toán phát hiện nhất quán danh tính.

## **NỘI DUNG VÀ PHƯƠNG PHÁP**

### **Nội dung 1: Nghiên cứu, khảo sát các hướng tiếp cận hiện có cho bài toán nhận dạng Deepfake**

- Tìm hiểu và nghiên cứu nhằm cung cấp cái nhìn tổng quát về các kết quả nghiên cứu của chủ đề tạo Deepfake, phát hiện Deepfake cũng như chống lại phát hiện Deepfake với các tài liệu được chọn lọc từ Google Scholar và khảo sát cẩn thận, đánh giá ưu và nhược điểm của các phương pháp hiện có, từ đó đề xuất hướng tiếp cận mong muốn

- Deepfake có thể được chia thành 4 loại chính là: entire face synthesis, attribute manipulation, identity swap, và expression swap. Trong đó hoán đổi danh tính (identity swap) xếp hạng cao nhất về mức độ phổ biến và nguy hiểm [3].
- Khảo sát các kết quả trước đó cho bài toán deepfake detection, kết quả hầu hết các phương pháp hiện có nhằm mục đích phân biệt hình ảnh giả bằng cách khai thác kết cấu cấp thấp và tìm kiếm các thao tác đối tượng bên dưới [3, 5]. Trong khi triển khai các kỹ thuật này trong các sản phẩm trong thế giới thực, chúng tôi quan sát thấy hai vấn đề phổ biến:
  - Phát hiện deepfake thường được thực hiện trên các video bị nghi ngờ và các khung hình video bị giảm chất lượng hình ảnh, chẳng hạn như thay đổi tỷ lệ hình ảnh, nhiễu và chuyển đổi codec video;
  - Khi deepfake được tạo giống như ảnh chân thực một cách thuyết phục, dấu vết giả mạo ở cấp độ thấp trở nên rất khó phát hiện. Những vấn đề này làm cho khả năng phát hiện deepfake không ổn định với đầu vào video. Chúng tôi muốn phát hiện deepfake mạnh mẽ hơn đáng kể bằng cách sử dụng nhiều thông tin nhận dạng có ý nghĩa về mặt ngữ nghĩa.

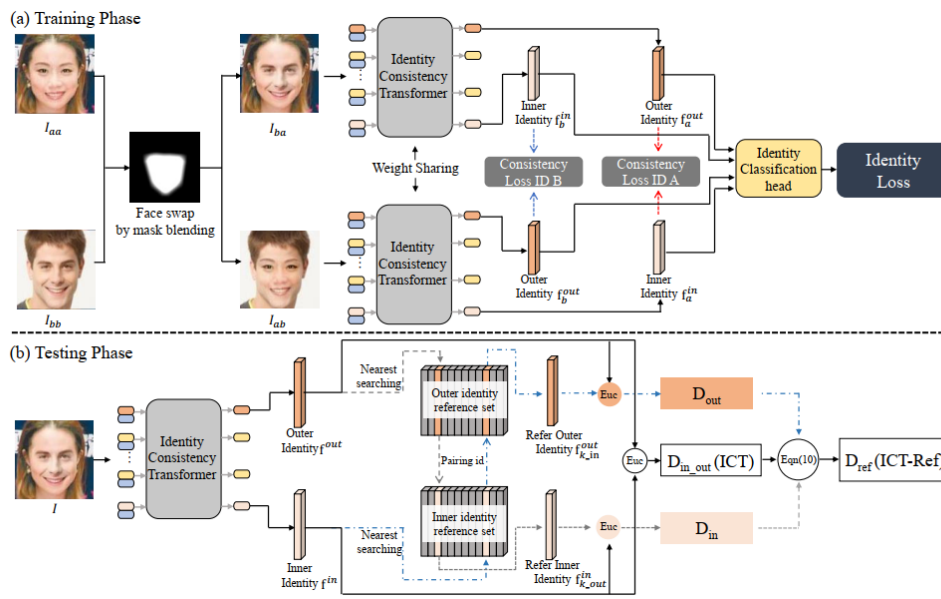
**Nội dung 2: Xây dựng mô hình phát hiện giả mạo khuôn mặt có tên là Identity Consistency Transformer (ICT) dựa trên thông tin ngữ nghĩa cấp cao.**

- Sử dụng kiến trúc **Transformer**: Mô hình ICT sử dụng kiến trúc **Transformer**, đây là một loại mạng thần kinh thường được sử dụng trong các tác vụ xử lý ngôn ngữ tự nhiên. Kiến trúc này chưa được sử dụng rộng rãi trong các nhiệm vụ thị giác máy tính và nghiên cứu cho



thấy rằng nó có thể hiệu quả trong việc học và lưu giữ các đặc điểm danh tính.

- Tập trung vào việc phát triển danh tính: Mô hình ICT được thiết kế để tập trung vào việc duy trì các đặc điểm nhận dạng của người nổi tiếng, thay vì chỉ phát hiện các điểm bất thường về hình ảnh trong video. Bằng cách học cách mã hóa và duy trì các đặc điểm nhận dạng của người nổi tiếng, mô hình có thể tạo các video tổng hợp duy trì các đặc điểm nhận dạng giống nhau trong khi thay đổi nội dung hình ảnh.



Hình 2: Minh họa giai đoạn (a) Training và (b) Testing của mô hình ICT

- Giai đoạn đào tạo: Giai đoạn đào tạo của mô hình ICT được thể hiện trong Hình 2a. Dữ liệu đào tạo bao gồm các video thực của người nổi tiếng và các video tổng hợp do mô hình ICT tạo ra. Các video thực được đưa vào bộ mã hóa (encoder), tạo ra một chuỗi các feature maps cho từng khung hình (frame). Các feature maps này sau đó được đưa vào bộ mã hóa Transformer (decoder Transformer), bộ mã hóa này sẽ học cách mã hóa các đặc điểm nhận dạng của người nổi tiếng. Các tính năng được

mã hóa sau đó được đưa vào bộ giải mã Transformer (decoder), bộ giải mã này sẽ tạo ra một chuỗi feature maps mới. Sau đó, các feature maps được đưa vào decoder, decoder này sẽ tạo ra một khung hình video tổng hợp cho từng bản đồ tính năng. Sau đó, các khung hình video tổng hợp được đưa vào discriminator để học cách phân biệt giữa các khung hình thực và tổng hợp. Hàm loss kết hợp adversarial loss, identity consistency loss, và perceptual loss để đào tạo generator và discriminator.

**Nội dung 3: Nghiên cứu, thực nghiệm và đề xuất cải tiến thuật toán Transformer hiện có để kiểm chứng hiệu quả với module bài toán phân loại khuôn mặt.**

- Nghiên cứu và thử nghiệm các thước đo tính nhất quán nhận dạng khác nhau vào thuật toán Transformer (đây là các kỹ thuật chính trong việc cải thiện chuẩn hóa nhằm tạo ra phiên bản ICT), so sánh và đánh giá các mô hình ứng với từng trường hợp áp dụng các kỹ thuật nhằm tìm ra mô hình phù hợp nhất và đạt được tối thiểu một trong ba mục tiêu đã đặt ra.
- Nghiên cứu thuật toán Transformer truyền thống trong bài toán deepfake detection từ đó phát triển và huấn luyện mô hình ICT trên tập dữ liệu MS-Celeb-1M. Cho phép mô hình tìm hiểu các đặc điểm nhận dạng qua nhiều tư thế, điều kiện ánh sáng và nét mặt, từ đó tạo các video tổng hợp duy trì các đặc điểm nhận dạng của người nổi tiếng trong các tình huống khác nhau.
- Đánh giá hiệu quả của mô hình ICT trên nhiều thang đo: AUC, saliency map,... và so sánh với các mô hình State-of-the-art cũng như mô phỏng kịch bản áp dụng ICT trong thế giới thực.
- Nghiên cứu các kỹ thuật tăng cường dữ liệu (Data Augmentation): biến

dạng mask, hiệu chỉnh màu sắc để hỗ trợ cho việc xây dựng bộ dữ liệu huấn luyện.

- Xây dựng chương trình ứng dụng minh họa bằng code Python.

## **KẾT QUẢ MONG ĐỢI**

- Tạo ra được mô hình ICT cho bài toán phát hiện các hình ảnh khuôn mặt giả mạo, tập trung vào ngữ nghĩa cấp cao và được tăng cường bằng cách tận dụng thông tin danh tính bổ sung từ những người nổi tiếng.
- Báo cáo phương pháp và kỹ thuật của mô hình ICT đã phát triển, kết quả thực nghiệm, đánh giá
- Cải thiện đáng kể khả năng phát hiện deepfake và bảo vệ quyền riêng tư của người nổi tiếng, cả hai đều là những vấn đề quan trọng trong bối cảnh kỹ thuật số hiện tại. Hy vọng rằng giải pháp của chúng tôi có thể khuyến khích nhiều kết quả nghiên cứu hơn trong công việc phát hiện giả mạo khuôn mặt tập trung vào điều tra sự không nhất quán trong ngữ nghĩa cấp cao trong công cuộc ngăn chặn sự lan truyền nội dung bị thao túng trên internet.
- Chúng tôi dự kiến công bố:
  - 01 bài báo hội nghị quốc tế thuộc danh mục CVPR.
  - 01 bài tạp chí Q3.
- Một chương trình demo để trực quan hóa nghiên cứu.

## **TÀI LIỆU THAM KHẢO** (*Định dạng DBLP*)

- [1] Ali Raza, Kashif Munir and Mubarak Almutairi. “A Novel Deep Learning Approach for Deepfake Image Detection”.
- [2]. Maryam Taeb and Hongmei Chi. “Comparison of Deepfake Detection Techniques through Deep Learning”.

[3]. “Countering Malicious DeepFakes: Survey, Battleground, and Horizon”.

[4] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis, 2018. 1, 2.

[5]. Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network, 2018. 1, 2, 6.