# THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/Y9KFP5WA38s*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/cuongnh20-uit/CS2205.SEP2025-face_x_ray/blob/main/C%C6%B0%E1%BB%9Dng%20Nguy%E1%BB%85n%20H%E1%BB%93ng%20-%20CS2205.SEP2025.DeCuong.FinalReport.Template.Slide.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Nguyễn Hồng Cường
- MSSV: 250202005

- Lớp: CS2205.SEP2025
- Tự đánh giá (điểm tổng kết môn): 8.5/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 3
- Số câu hỏi QT của cả nhóm: 15
- Link Github: https://github.com/cuongnh20-uit/CS2205.SEP2025-face_x_ray

**TÊN ĐỀ TÀI (IN HOA)**

NGHIÊN CỨU VÀ PHÁT TRIỂN HỆ THỐNG PHÁT HIỆN GIẢ MẠO KHUÔN MẶT DỰA TRÊN PHÂN TÍCH ĐƯỜNG BIÊN PHA TRỘN (FACE X-RAY).

**TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**

FACE X-RAY: A GENERALIZABLE DEEPFAKE DETECTION FRAMEWORK BASED ON BLENDING BOUNDARY ANALYSIS.

**TÓM TẮT** *(Tối đa 400 từ)*

Trong bối cảnh bùng nổ của các công nghệ sinh ảnh giả mạo (Deepfake), khả năng xác thực danh tính trong an toàn thông tin đang đối mặt với những thách thức nghiêm trọng. Các phương pháp phát hiện hiện có chủ yếu dựa vào việc học các hiện vật (artifacts) cụ thể của từng thuật toán tạo giả, dẫn đến khả năng tổng quát hóa kém khi đối mặt với các dạng tấn công chưa biết (unseen attacks). Nguyên nhân chính là sự phụ thuộc quá mức vào các đặc trưng ngữ nghĩa mức cao thay vì các dấu hiệu giả mạo cơ bản. Để khắc phục hạn chế này, đề cương đề xuất nghiên cứu áp dụng phương pháp Face X-ray, tập trung vào việc phát hiện đường biên pha trộn (blending boundary) – một bước kỹ thuật cơ bản tồn tại trong hầu hết các quy trình tráo đổi khuôn mặt hiện nay. Hệ thống được xây dựng dựa trên mạng nơ-ron tích chập (CNN) để dự đoán ranh giới pha trộn và phân loại ảnh thật/giả mà không cần tri thức trước về thuật toán tạo Deepfake cụ thể. Đề tài sử dụng bộ dữ liệu FaceForensics++ để huấn luyện và đánh giá, hướng tới mục tiêu nâng cao độ chính xác và khả năng chống chịu trước các biến thể tấn công mới, góp phần bảo vệ tính toàn vẹn của dữ liệu đa phương tiện.

**GIỚI THIỆU** *(Tối đa 1 trang A4)*

Sự phát triển nhanh chóng của các kỹ thuật Generative Adversarial Networks (GANs)

và Autoencoders đã cho phép tạo ra các video giả mạo khuôn mặt (Deepfake) với độ chân thực cao, khó phân biệt bằng mắt thường. Trong lĩnh vực an toàn thông tin, điều này gây ra các nguy cơ tiềm ẩn về lừa đảo trực tuyến, giả mạo xác thực sinh trắc học và thao túng thông tin truyền thông. Mặc dù nhiều giải pháp phát hiện đã được đề xuất, phần lớn hoạt động như một bài toán phân loại nhị phân thông thường, học các đặc trưng lỗi ảnh cụ thể của một số công cụ tạo Deepfake nhất định.

Tuy nhiên, thực trạng cho thấy các công cụ tạo Deepfake liên tục được cập nhật và cải tiến. Các mô hình phát hiện truyền thống thường gặp hiện tượng "quá khớp" (overfitting), dẫn đến hiệu suất giảm sút nghiêm trọng khi kiểm thử trên các video được tạo bởi các thuật toán chưa từng gặp trong tập huấn luyện. Khoảng trống nghiên cứu nằm ở việc thiếu các phương pháp tiếp cận dựa trên các đặc trưng bất biến hoặc cốt lõi của quy trình giả mạo thay vì các lỗi ảnh bề mặt.

Một quan sát quan trọng trong quy trình tạo Deepfake là bước pha trộn (blending): khuôn mặt giả được ghép đè lên ảnh gốc. Quá trình này luôn để lại các dấu vết thống kê tại vùng biên giới giữa hai nguồn ảnh khác nhau. Do đó, việc nghiên cứu một giải pháp tập trung vào phát hiện đường biên pha trộn (Face X-ray) có ý nghĩa thực tiễn cao, hứa hẹn mang lại khả năng tổng quát hóa tốt hơn (Generalization), độc lập với thuật toán tạo giả cụ thể. Đây là cơ sở để đề tài tập trung nghiên cứu và hiện thực hóa giải pháp này.

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

1. Nghiên cứu và hệ thống hóa các kỹ thuật tạo và phát hiện Deepfake, đặc biệt là các phương pháp dựa trên phân tích tính không nhất quán của ảnh và đường biên pha trộn.
2. Xây dựng và huấn luyện mô hình Face X-ray dựa trên kiến trúc HRNet (High-Resolution Network) để dự đoán vùng biên giới pha trộn và phân loại video thật/giả, sử dụng bộ dữ liệu FaceForensics++.
3. Đánh giá hiệu quả của mô hình thông qua các chỉ số độ chính xác (Accuracy), diện tích dưới đường cong (AUC) và khả năng tổng quát hóa trên các tập dữ liệu chéo (cross-dataset evaluation).

## NỘI DUNG VÀ PHƯƠNG PHÁP

### 1. Nội dung

Đề tài tập trung giải quyết bài toán phát hiện Deepfake dưới góc độ Generalization (tính tổng quát). Nội dung nghiên cứu bao gồm việc phân tích quy trình tạo ảnh giả mạo để xác định bước "pha trộn" là bước chung nhất của hầu hết các thuật toán. Từ đó, xây dựng cơ chế tự tạo dữ liệu huấn luyện bằng cách tráo đổi vùng mặt và tạo ra các biên giới pha trộn nhân tạo, giúp mô hình học được đặc trưng ranh giới thay vì nội dung khuôn mặt. Cuối cùng, hệ thống được thiết kế để đưa ra quyết định dựa trên sự tồn tại của các đường biên này.

### 2. Phương pháp

Nghiên cứu áp dụng phương pháp thực nghiệm dựa trên mô hình học sâu (Deep Learning), cụ thể:

- **Chuẩn bị dữ liệu (Data Preparation):** Sử dụng bộ dữ liệu FaceForensics++ (bao gồm 4 phương pháp giả mạo: Deepfakes, Face2Face, FaceSwap,

NeuralTextures). Dữ liệu video được tách thành các frame và trích xuất vùng khuôn mặt. Một kỹ thuật quan trọng được áp dụng là **Dynamic Data Augmentation**: tạo ra các mẫu huấn luyện bằng cách pha trộn một khuôn mặt từ ảnh A sang ảnh B với các đường biên Gaussian mờ, tạo ra nhãn (Ground Truth) là chính các đường biên đó.

- **Kiến trúc mô hình:** Sử dụng mạng HRNet (High-Resolution Net) làm backbone để trích xuất đặc trưng. Mạng này có ưu điểm giữ được độ phân giải cao qua các tầng, rất phù hợp để phát hiện các chi tiết nhỏ như đường biên pha trộn. Đầu ra của mạng là một bản đồ xám (grayscale map) thể hiện xác suất pha trộn tại từng pixel và một đầu ra phân loại (Softmax) cho kết quả Thật/Giả.
- **Hàm mất mát (Loss Function):** Sử dụng kết hợp Cross-Entropy Loss cho bài toán phân loại và một hàm loss hồi quy (ví dụ Mean Squared Error) để so sánh bản đồ biên dự đoán với bản đồ biên thực tế.
- **Quy trình đánh giá:** Mô hình được huấn luyện trên tập train của FaceForensics++ và đánh giá trên tập test. Ngoài ra, để kiểm tra tính tổng quát, mô hình sẽ được kiểm thử trên các dataset chưa từng nhìn thấy (ví dụ: DeepFake Detection Challenge - DFDC) để chứng minh hiệu quả so với các phương pháp truyền thống như XceptionNet.

## KẾT QUẢ MONG ĐỢI

1. Một bộ mã nguồn hoàn chỉnh thực thi mô hình Face X-ray, cho phép huấn luyện lại và kiểm thử trên các video đầu vào tùy ý.
2. Mô hình đạt độ chính xác (Accuracy) trên 90% đối với tập dữ liệu FaceForensics++ (chất lượng cao) và duy trì được chỉ số AUC ổn định (>70%) khi đánh giá trên các loại Deepfake chưa từng gặp trong quá trình huấn luyện.
3. Báo cáo chi tiết so sánh hiệu năng giữa phương pháp Face X-ray và các baseline (như XceptionNet, MesoNet), minh chứng được tính ưu việt trong khả năng tổng quát hóa.

4. Ứng dụng demo (Web hoặc Script) cho phép người dùng tải lên video và nhận kết quả phân tích xác suất giả mạo kèm theo hình ảnh trực quan hóa vùng biên pha trộn.

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

[1] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. "Face X-ray for More General Face Forgery Detection." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5001-5010, 2020.

[2] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. "FaceForensics++: Learning to Detect Manipulated Facial Images." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1-11, 2019.

[3] Yuezun Li and Siwei Lyu. "Exposing DeepFake Videos By Detecting Face Warping Artifacts." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

# ĐỀ CƯƠNG NGHIÊN CỨU