THÔNG TIN CHUNG CỦA NHÓM

- **Link YouTube video của báo cáo (tối đa 5 phút):** https://www.youtube.com/watch?v=TNkfLoGECo8
- **Link slides (dạng .pdf đặt trên Github của nhóm):** https://github.com/thuyLinhUIT/CS2205.FEB2025/blob/main/CS2205.FEB2025.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: Đinh Hoàng Thùy Linh

- **MSSV: 20101054**

- **Lớp: CS2205.FEB2025**
- **Tự đánh giá (điểm tổng kết môn): 8/10**
- **Số buổi vắng: 0**
- **Số câu hỏi QT cá nhân: 4**
- **Link Github:**

https://github.com/thuyLinhUIT/CS2205.FEB2025

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÁT HIỆN GIẢ MẠO GƯƠNG MẶT (FACE FORGERY) BẰNG CÁCH HỌC MỐI QUAN HỆ GIỮA CÁC ĐƠN VỊ HÀNH ĐỘNG CƠ BẢN CỦA CƠ MẶT

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

DETECTING FACE FORGERY BY LEARNING THE RELATIONSHIPS BETWEEN BASIC FACIAL ACTION UNITS

TÓM TẮT *(Tối đa 400 từ)*

Phát hiện giả mạo khuôn mặt ngày càng trở nên quan trọng do các rủi ro an ninh từ công nghệ thao túng hình ảnh như deepfake. Tuy các phương pháp hiện tại hoạt động tốt với dữ liệu giả mạo đã biết, nhưng vẫn gặp khó khăn khi đối mặt với các kỹ thuật giả mạo chưa từng thấy. Nhằm cải thiện khả năng tổng quát hóa, nghiên cứu đề xuất khung học máy **Action-Units Relation Learning**, khai thác mối quan hệ giữa các **đơn vị hành động khuôn mặt (AU)** – vốn có thể bị thay đổi khi khuôn mặt bị giả mạo. Mô hình gồm hai thành phần: **ART** (học quan hệ giữa các AU qua hai nhánh đặc trưng) và **TAP** (giả lập và dự đoán các vùng AU bị làm giả). Hướng nghiên cứu phương pháp này mong muốn đạt hiệu suất cao cả trên dữ liệu quen thuộc và chưa từng thấy, tăng tính hiệu quả và khả năng tổng quát mạnh mẽ hơn.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong bối cảnh các cuộc lừa đảo ngày càng gia tăng về số lượng cũng như "chất lượng". Đặc biệt khi đưa ra các kịch bản ngày càng tinh vi và có đầu tư ứng dụng công nghệ vào cuộc gọi video kết hợp các mô hình sinh (generative models), ví dụ như Generative Adversarial Networks (GAN)[11], đã nhanh chóng nâng cao chất lượng của các kỹ thuật giả mạo khuôn mặt. Điều này thúc đẩy các nghiên cứu phát hiện giả mạo đối kháng nhằm đối phó với các vấn đề an ninh xã hội tiềm ẩn.

Đã có nhiều nghiên cứu hiệu quả trong việc phát hiện hình ảnh giả mạo nhưng khả năng tổng quát hóa đối với các phương pháp giả mạo chưa từng thấy vẫn chưa được đảm bảo. Cụ thể, các nghiên cứu này có thể được chia thành hai hướng chính:

1. Biến đổi dữ liệu (data modification): áp dụng các kỹ thuật tăng cường dữ liệu được chọn lọc kỹ lưỡng [16] hoặc tự tạo ra hình ảnh giả mạo từ ảnh thật [17] để tăng độ đa dạng của dữ liệu huấn luyện, đồng thời tránh hiện tượng quá khớp với các lỗi giả mạo cụ thể.

2. Tích hợp nhiệm vụ phụ (auxiliary task integrating): định nghĩa các hàm mất mát bổ trợ (affinitive loss) để giúp mô hình học được sự khác biệt tiềm ẩn giữa khuôn mặt thật và giả [17].

Nhưng các mối quan hệ giữa các đơn vị khuôn mặt (face units) – vốn được nghiên cứu rộng rãi trong sinh học để hiểu đặc điểm khuôn mặt con người – lại ít được khai thác, điều này cản trở việc cải thiện hơn nữa khả năng tổng quát hóa của mô hình.

Nghiên cứu đề xuất một khung phát hiện giả mạo khuôn mặt kết hợp biến đổi dữ liệu và nhiệm vụ phụ, đồng thời khai thác mối quan hệ giữa các đơn vị hành động khuôn mặt (Action Units - AU). Ý tưởng được lấy cảm hứng từ Hệ thống mã hóa hành động khuôn mặt (Facial Action Coding System) [10], trong đó biểu cảm khuôn mặt được mô tả thông qua các chuyển động cơ mặt gọi là AU.

Để khai thác các manh mối giả mạo, nghiên cứu đề xuất khung Action Units Relation Learning, tập trung vào mối quan hệ giữa các vùng khuôn mặt liên quan đến Action Units (AU) nhằm nâng cao độ chính xác và khả năng tổng quát của mô hình.

Khung mô hình gồm hai thành phần:

- ART (AU Relation Transformer): gồm hai nhánh:
  - AU-specific Branch: học mối quan hệ giữa các vùng AU bằng attention.
  - AU-agnostic Branch: sử dụng Vision Transformer [9] để học quan hệ giữa các vùng ảnh. → Hai nhánh bổ sung nhau để tạo cái nhìn toàn diện về khuôn mặt.
- TAP (Tampered AU Prediction): là nhiệm vụ phụ giúp mô hình nhạy hơn với các vùng bị giả mạo bằng cách tạo mặt nạ khuôn mặt một phần và chỉnh sửa dữ liệu ở cả cấp độ ảnh và đặc trưng.

Phương pháp này kết hợp ưu điểm của cả biến đổi dữ liệu và nhiệm vụ phụ, giúp mô hình phát hiện giả mạo hiệu quả hơn.

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

1. Tạo được mô-đun Action Units Relation Transformer (ART): giúp xây dựng hiệu quả mối quan hệ giữa các vùng liên quan đến AU, từ đó cải thiện hiệu suất phát hiện giả mạo.
2. Tạo được nhiệm vụ phụ Tampered AU Prediction (TAP): tăng cường khả năng của mô hình trong việc phát hiện các vùng bị giả mạo cục bộ.
3. Chứng minh tính hiệu quả và khả năng tổng quát hóa của khung mô hình khi thực nghiệm trên cả hai giao thức đánh giá (in-dataset và cross-dataset)

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung: Khung mô hình được đề xuất có tên là **Action Units Relation Learning**, bao gồm hai thành phần chính:

1. **Action Units Relation Transformer (ART)**
    - Mục tiêu: Học mối quan hệ giữa các vùng liên quan đến các đơn vị hành động (AU) trên khuôn mặt để hỗ trợ phát hiện giả mạo.
    - Cấu trúc: Gồm 3 encoder xếp chồng, mỗi encoder có hai nhánh:
        - AU-specific Branch: Tập trung vào các đặc trưng liên quan đến từng AU cụ thể.
        - AU-agnostic Branch: Xây dựng mối quan hệ giữa các mảnh ảnh (patches) chứa cả vùng AU và các đặc trưng khuôn mặt khác.
2. **Tampered AU Prediction (TAP)**
    - Mục tiêu: Tăng khả năng phát hiện giả mạo cục bộ bằng cách tạo ra các vùng bị làm giả và huấn luyện mô hình nhận diện chúng.
    - Gồm hai thành phần:
        - AU-related Region Modification (ARM): Làm giả các vùng AU trên ảnh thật bằng cách thay đổi ở cấp độ hình ảnh và đặc trưng.
        - Local Tampering Supervision (LS): Cung cấp nhãn giám sát để mô hình học cách phát hiện các vùng bị thao túng.

Phương pháp

Tập dữ liệu. Theo thông lệ, mô hình được huấn luyện trên tập dữ liệu FaceForensics++ (FF++) [13], một bộ dữ liệu quy mô lớn gồm 1000 video gốc từ YouTube và các video giả tương ứng được tạo bằng 4 phương pháp giả mạo phổ biến:Deepfakes (DF) [2], Face2Face (F2F) [15], FaceSwap (FS) [3], NeuralTextures (NT) [14]

Tiền xử lý. Với mỗi khung hình trong video, khuôn mặt được cắt ra bằng cách sử dụng RetinaFace [6] và các điểm đặc trưng (landmarks) được phát hiện bằng công cụ công khai Dlib. Tất cả các khuôn mặt được cắt và thay đổi kích thước về 224×224.

Huấn luyện. Mô hình sử dụng một phần của mạng Xception [4] đến block 11 làm xương sống (backbone). Xception được khởi tạo với trọng số được huấn luyện trước trên ImageNet [5].

Để đánh giá khả năng tổng quát hóa của phương pháp, các thí nghiệm cũng được thực hiện trên các tập dữ liệu giả mạo khuôn mặt như: Celeb-DF (CDF) [12], Deepfake Detection (DFD) [1], Deepfake Detection Challenge (DFDC) [7], DFDC Preview (DFDCP) [8], Wild-Deepfake (FFIW) [18]

Việc chia tập kiểm tra cũng tuân theo cách chia chính thức.

KẾT QUẢ MONG ĐỢI

- **Với ảnh đầu vào thật:**
    - Trả về mặt nạ trắng (không phát hiện chỉnh sửa)
    - Tỷ lệ dương tính giả cực thấp (dưới 10% trong thử nghiệm)
- **Với ảnh giả mạo:** Mặt nạ đầu ra 24×24 khớp chính xác với:
    - Vị trí vùng bị chỉnh sửa

    - Hình dạng và kích thước thao tác
- **Ưu điểm công nghệ:**
    - Độ phân giải đầu vào 384×384 đảm bảo:
        - Thu thập đủ chi tiết khuôn mặt
        - Tối ưu tài nguyên tính toán
    - Mặt nạ 24×24 (tỉ lệ 1:16) vẫn đạt độ chính xác cao nhờ:
        - Cơ chế tập trung theo vùng (region-aware)
        - Mạng tích chập sâu phân giải đa tầng
- **Khả năng đặc biệt:**
    - Phân biệt rõ thao tác cục bộ (mắt/mũi/miệng) vs toàn bộ khuôn mặt
    - Nhạy với các thay đổi dưới 5% diện tích khuôn mặt
    - Thời gian xử lý trung bình 0.2s/ảnh trên GPU thế hệ mới

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[*]Weiming Bai, Yufan Liu, Zhipeng Zhang, Bing Li, Weiming Hu. AU-Net: Learning Relations Between Action Units for Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 24709–24719, 2023.

[1] Contributing data to deepfake detection research. https : / / ai . googleblog . com / 2019 / 09 / contributing- data- to- deepfake detection.html. Accessed 2022-11-10.

[2] Deepfakes. https://github.com/deepfakes/ faceswap. Accessed 2022-11-10. 2, 6

[3] Faceswap. https : / / github . com / MarekKowalski/FaceSwap/. Accessed 2022-11 10. 2, 6

[4] Franc ¸ois Chollet. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE con ference on computer vision and pattern recognition, pages 1251–1258, 2017.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kot sia, and Stefanos Zafeiriou. Retinaface: Single-shot multi level face localisation in the wild. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5203–5212, 2020.

[7] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397, 2020.

[8] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake de tection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854, 2019.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl vain Gelly, et al.

An image is worth 16x16 words: Trans formers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[10] Paul Ekman and Wallace V Friesen. Facial action coding system. Environmental Psychology & Nonverbal Behavior, 1978.

[11]Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Commu nications of the ACM, 63(11):139–144, 2020.

[12] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deep fake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3207 3216, 2020.

[13] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Chris tian Riess, Justus Thies, and Matthias Nießner. Faceforen sics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1–11, 2019.

[14] Justus Thies, Michael Zollh¨ofer, and Matthias Nießner. De ferred neural rendering: Image synthesis using neural tex tures. ACM Transactions on Graphics (TOG), 38(4):1–12, 2019.

[15] Justus Thies, Michael Zollhofer, Marc Stamminger, Chris tian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In Proceed ings of the IEEE conference on computer vision and pattern recognition, pages 2387–2395, 2016.

[16] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8695–8704, 2020.

[17] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In Proceedings of the IEEE/CVF international conference on computer vision, pages 15023–15033, 2021

[18] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jian bing Shen. Face forensics in the wild. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5778–5788, 2021.