p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

MỘT PHƯƠNG PHÁP TIẾP CẬN MỚI ĐỂ PHÁT HIỆN VIDEO DEEPFAKE

Lê Văn Hào[1], Trần Doãn Minh[1], Trịnh Thị Hợp[1], Lê Diệu Linh[1]

TÓM TẮT

Trong bối cảnh kỹ thuật tạo video bằng công nghệ Deepfake ngày càng tinh vi, việc phát triển các phương pháp phát hiện tự động chính xác và đáng tin cậy trở nên cấp thiết. Nghiên cứu này đề xuất mô hình ViT-DFNet, tập trung vào phát hiện hiệu quả các trường hợp hoán đổi khuôn mặt trong video. Kết quả thực nghiệm cho thấy ViT-DFNet đạt độ chính xác và khả năng tổng quát cao hơn so với các kiến trúc mạng nơ-ron tích chập truyền thống. Đồng thời, nghiên cứu cũng phân tích khả năng diễn giải của mô hình thông qua bản đồ nhiệt để làm rõ những khu vực được hệ thống tập trung khi phát hiện dấu hiệu giả mạo.

Từ khóa: *Phát hiện Deepfake, học sâu, phân loại ảnh.*

DOI: *https://doi.org/10.70117/hdujs.84.2.2026.1034*

1. ĐẶT VẤN ĐỀ

Deepfake, thuật ngữ bắt nguồn từ sự kết hợp giữa 'học sâu' (Deep Learning) và 'giả mạo' (Fake), dùng để mô tả kỹ thuật tổng hợp nội dung số (bao gồm hình ảnh, video, và âm thanh) để tạo ra các nội dung giả mạo gần với video thực [1]. Sự phổ biến nhanh chóng của công nghệ này, đặc biệt là hình thức hoán đổi khuôn mặt trong các hình ảnh hoặc video, đang đặt ra những mối đe dọa nghiêm trọng tới quyền riêng tư, an ninh xã hội, dân chủ và niềm tin vào các phương tiện truyền thông số [2][4]. Kẻ xấu lợi dụng Deepfake để lan truyền thông tin sai lệch, gây bất ổn chính trị, hoặc gian lận tài chính [3]. Do đó, trong 5 năm trở lại đây, chủ đề nghiên cứu phát hiện video giả mạo được tạo ra bởi công nghệ Deepfake đang thu hút đáng kể sự quan tâm của cộng đồng nghiên cứu [1-4].

Để giải quyết những thách thức trên, trong bài báo này, chúng tôi đề xuất một mô hình ViT-DFNet, dựa trên kiến trúc thế hệ mới Vision Transformer - ViT [12]. Mô hình ViT vốn nổi tiếng với khả năng nắm bắt mối quan hệ phụ thuộc toàn cục trong chuỗi dữ liệu, được kỳ vọng sẽ vượt qua giới hạn của các mô hình mạng nơ-ron tích chập (CNN) truyền thống, vốn chủ yếu tập trung vào các đặc trưng cục bộ.

Bài báo được cấu trúc các phần như sau. Phần đầu cung cấp những khái niệm cơ bản về công nghệ Deepfake. Kế đến, phần hai chúng tôi trình bày về các nghiên cứu liên quan và đề xuất giải pháp cho phép phát hiện sự hoán đổi khuôn mặt trong các video giả mạo được tạo ra từ công nghệ Deepfake. Các kết quả nghiên cứu được trình bày ở phần 3. Cuối cùng là các kết luận và một số hướng nghiên cứu trong tương lai.

[1] *Khoa Kỹ thuật, Công nghệ và Truyền thông, Trường Đại học Hồng Đức; Email: levanhao@hdu.edu.vn*

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Phương pháp tiên tiến hiện nay

Về bản chất bài toán phát hiện video giả mạo được xem là một vấn đề phân loại nhị phân [2][5]. Trong đó, các phương pháp chủ yếu được phân loại dựa trên miền không gian, miền thời gian, và miền tần số [2][6]. Các phương pháp trong miền không gian khai thác các giả mạo hình ảnh, đặc biệt là các biến dạng khuôn mặt (face warping) được tạo ra trong quá trình thay thế khuôn mặt [8], sử dụng các mô hình mạng nơ-ron tích chập phổ biến như VGG16, ResNet50 [2][4-6], hoặc các kiến trúc mạng chuyên biệt như MesoNet [9]. Hướng nghiên cứu về thời gian tập trung vào không nhất quán thời gian giữa các khung hình [7], ví dụ điển hình là sự thiếu vắng hành vi nháy mắt chân thực [7][10], hoặc sự mâu thuẫn về tư thế đầu [5], thường sử dụng các mạng nơ-ron hồi quy (Recurrent Neural Networks - RNN), hoặc phiên bản cải tiến trong các mạng Long Short-Term Memory (LSTM) kết hợp với các mạng nơ-ron tích chập CNN để trích xuất các đặc trưng không gian - thời gian [7][11]. Ngoài ra, việc phân tích trong miền tần số cũng được áp dụng để tìm kiếm dấu vết giả mạo tinh vi [7]. Tuy nhiên, thách thức lớn nhất đối với các mô hình hiện tại là khả năng tổng quát hóa đối với các nguồn giả mạo không quen thuộc hoặc các kỹ thuật Deepfake thế hệ mới.

Tại Việt Nam, nghiên cứu về phát hiện Deepfake cũng đã bắt đầu phát triển, tập trung vào việc áp dụng phương pháp học chuyển giao để giải quyết bài toán, thử nghiệm và đánh giá hiệu suất của các mô hình mạng nơ-ron tích chập như XceptionNet, ResNet50, VGG19, và DenseNet trên các tập dữ liệu công khai [1][3]. Nhìn chung, các nghiên cứu này đang sử dụng những kiến trúc mạng nơ-ron tích chập truyền thống, sẽ là những hạn chế trong việc đánh giá khả năng tổng quát của chúng đối với những video Deepfake thế hệ mới.

2.2. Giải pháp đề xuất

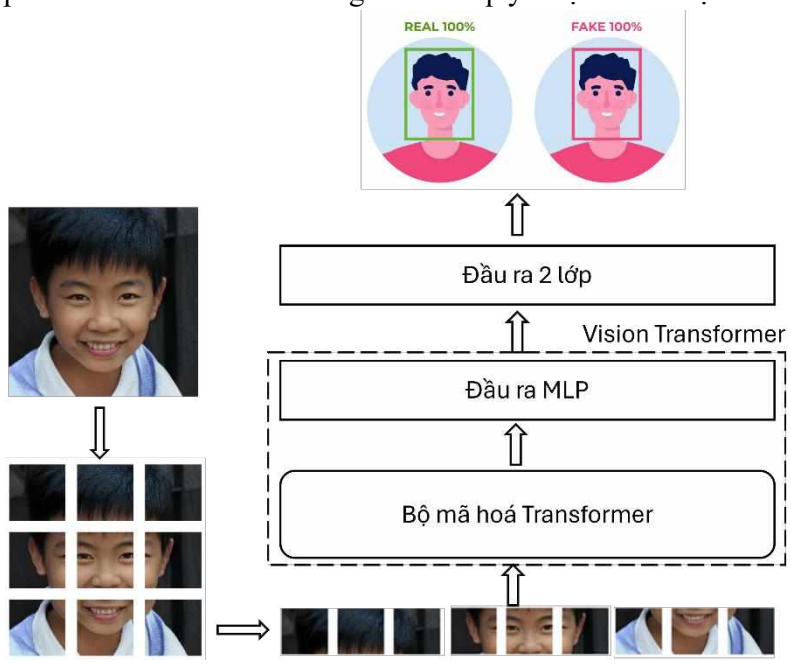
Dựa trên các kết quả nghiên cứu hiện tại, chúng tôi nhận thấy rằng các giải pháp dựa trên phương pháp phân loại từng khung hình (frame-by-frame) đã được chứng minh là hiệu quả, đặc biệt là phát hiện các thao tác trong hoán đổi khuôn mặt của các video tạo bằng công nghệ Deepfake [4][7][8]. Bên cạnh đó, các mô hình mạng nơ-ron tích chập như XceptionNet, ResNet50, và DenseNet121 đã được sử dụng rộng rãi để phát hiện các tạo tác này [2][4-6].

Trong bài báo này, chúng tôi đề xuất một mô hình phát hiện sự hoán đổi khuôn mặt trong các video tạo bởi công nghệ Deepfake, có tên là ViT-DFNet. Kiến trúc mô hình ViT-DFNet được minh họa trong Hình 1. Mô hình này được thiết kế dựa trên mô hình có tên Vision Transformer (ViT) [12]. Kiến trúc ViT dựa trên cơ chế tự chú ý (self-attention), một mô hình nổi tiếng với khả năng nắm bắt mối quan hệ phụ thuộc toàn cục trong chuỗi dữ liệu. Những đặc điểm này phù hợp với bản chất của nội dung Deepfake, vốn chứa các bất thường phân tán và khó quan sát. Vì vậy, việc áp dụng ViT vào phát hiện Deepfake có cơ sở khoa học và hứa hẹn mang lại kết quả chính xác hơn.

Để kiểm chứng mô hình, chúng tôi phân loại từng khung hình của video dựa trên việc biểu diễn đặc trưng của từng vùng ảnh nhỏ kết hợp với mô hình Transformer để phát hiện hoán đổi khuôn mặt. Cụ thể, từ ảnh đầu vào, chúng tôi tiền xử lý và tách mặt ra thành lưới nhỏ rồi chuyển đổi mỗi ô nhỏ thành một chuỗi ảnh theo thứ tự không gian. Các chuỗi này được ánh xạ thành vec-tơ đặc trưng và cộng với vị trí để tạo thành đầu vào cho bộ mã hóa

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

Transformer. Bộ mã hóa sẽ học các tương tác không gian giữa các vùng ảnh và xuất ra một vec-tơ đặc trưng tổng hợp. Các vec-tơ đặc trưng này được đưa qua một khối MLP để tinh chỉnh đặc trưng, sau đó qua một tầng đầu ra 2 lớp để phân loại nhãn Thật / Giả (REAL / FAKE). Ở cấp video, chúng tôi áp dụng quy trình trên cho từng khung hình có mặt, sau đó thực hiện bỏ phiếu đa số trên cửa sổ thời gian để ra quyết định cho đoạn video.



Hình 1. Kiến trúc mô hình ViT-DFNet

Để đánh giá hiệu suất của mô hình đề xuất. Chúng tôi sử dụng ma trận nhầm lẫn, một công cụ phổ biến trong bài toán phân loại nhị phân. Từ ma trận này, các chỉ số quan trọng như độ chính xác (Precision), độ bao phủ (Recall), F_1 được báo cáo. Các kết quả này cũng được so sánh với hiệu suất của mô hình nơ-ron tích chập truyền thống để chứng minh khả năng mạnh mẽ của mô hình đề xuất.

3. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

3.1. Bộ dữ liệu

Nhiều bộ dữ liệu đã được xây dựng phục vụ nghiên cứu phát hiện các nội dung tạo ra bởi Deepfake như FaceForensics++ [13], Celeb-DF [15], DFDC [4], Kodf [10], 140K-DB [14], hay Openforensics [16], được liệt kê ở Bảng 1. Trong nghiên cứu này, chúng tôi sử dụng hai bộ dữ liệu 140K-DB [14] và Openforensics [16]. Bộ dữ liệu 140K-DB [14] được lưu trữ trên Kaggle[2], gồm khoảng 140 000 ảnh mặt, chia đều thành 70 000 ảnh thật (REAL) và 70 000 ảnh giả (FAKE) - được tổng hợp từ 1 000 000 khuôn mặt giả. Tương tự như vậy, bộ dữ liệu Openforensics [16], được công bố trên nền tảng Github[3], bao gồm 115 000 hình

[2] *Bộ dữ liệu 140K-DB : https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces*
[3] *Bộ dữ liệu Openforensics: https://github. com/ltnghia/openforensics*

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

ảnh thực tế với 334 000 khuôn mặt người. Hai bộ dữ liệu này là được thu thập từ nhiều nguồn khác nhau, đảm bảo sự đa dạng về giới tính, độ tuổi, góc chụp, và điều kiện ánh sáng. Ngoài ra, số lượng lớn và cân bằng giữa lớp thật và lớp giả giúp giảm thiểu vấn đề thiên lệch lớp - điều rất quan trọng cho bài toán phân loại nhị phân. Đặc biệt các hình ảnh trong bộ dữ liệu Openforensics còn có thể có nhiều hơn một khuôn mặt.

Bảng 1. Mô tả các bộ dữ liệu chính được dùng trong các nghiên cứu hiện nay

Bộ dữ liệu	Số lượng mẫu thật	Số lượng mẫu giả	Tổng số
FaceForensics++ [13]	1 000	4 000	5 000
Celeb-DF [15]	590	5 639	6 229
DFDC [4]	23 654	104 500	128 154
Kodf [10]	62 166	175 776	237 942
140K-DB [14]	70 000	70 000	140 000
Openforensics [16]	95 201	95 134	190 335

Ngoài ra, chúng tôi thực hiện đánh giá chéo (cross-evaluation) nhằm kiểm tra khả năng khái quát hóa của mô hình phát hiện giả mạo khuôn mặt. Cụ thể, mô hình được huấn luyện trên bộ dữ liệu OpenForensics, vốn chứa nhiều khuôn mặt trong môi trường thực tế phức tạp, sau đó được đánh giá trên bộ dữ liệu 140K-DB.

3.2. Thiết lập môi trường

Quá trình huấn luyện và kiểm thử mô hình được thực hiện trên máy trạm trang bị GPU NVIDIA RTX 2080 T1 (12GB VRAM), CPU Intel(R) Xeon(R) W-2133 i6, RAM 32GB, chạy trên Ubuntu 22.04 LTS. Mô hình được triển khai bằng ngôn ngữ lập trình Python với các thư viện PyTorch, OpenCV, NumPy, và scikit-learn.

Trong quá trình huấn luyện mô hình ViT-DFNet, chúng tôi sử dụng kỹ thuật học chuyển giao để tinh chỉnh các trọng số trong mô hình ViT đã tiền huấn luyện trên bộ dữ liệu ImageNet. Bên cạnh đó, các siêu tham số được lựa chọn dựa trên các tiêu chuẩn phổ biến trong những nghiên cứu trước đây về ViT và phát hiện video giả mạo bằng Deepfake. Toàn bộ ảnh đầu vào được chuẩn hóa và thay đổi kích thước về 224x224 pixel, sau đó chia thành các khối nhỏ kích thước 16x16 pixel, phù hợp với cấu trúc gốc của ViT. Quá trình huấn luyện sử dụng Learning rate = 1e-3, Batch size = 1 024, số Epoch = 50, cùng bộ tối ưu Adam với hệ số Weight decay = 5e-4. Ngoài ra, mô hình áp dụng kỹ thuật Dropout = 0.4 và Early Stopping dựa trên hiệu suất tập xác thực để tránh quá khớp.

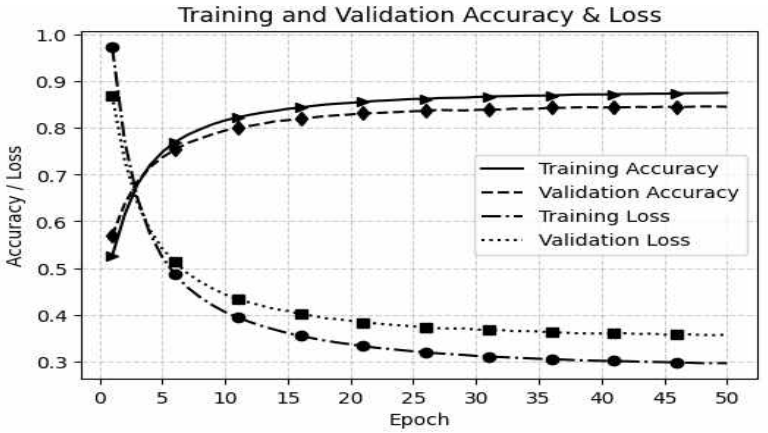
3.3. Kết quả

3.3.1. Kết quả quá trình huấn luyện

Hình 2 minh họa kết quả huấn luyện của mô hình ViT-DFNet trên bộ dữ liệu OpenForensics. Kết quả cho thấy mô hình đạt được sự hội tụ ổn định và hiệu quả sau khoảng 40-50 epoch. Cụ thể, độ chính xác huấn luyện (Training Accuracy) tăng đều đặn từ khoảng 0.6 ở giai đoạn đầu lên đến xấp xỉ 0.88 ở cuối quá trình. Độ chính xác kiểm định (Validation Accuracy) cũng có xu hướng tăng tương tự, đạt mức khoảng 0.85, chứng tỏ mô hình duy trì khả năng tổng quát tốt và không xảy ra hiện tượng quá khớp đáng kể.

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

Ngược lại, hàm mục tiêu huấn luyện (Training Loss) giảm mạnh từ gần 0.95 xuống dưới 0.3, trong khi hàm mục tiêu kiểm định (Validation Loss) cũng giảm dần và ổn định quanh mức 0.35 ở các epoch cuối. Sự song song của hai đường này ở giai đoạn cuối phản ánh quá trình tối ưu hóa diễn ra hiệu quả và mô hình đã học được các đặc trưng cốt lõi của dữ liệu.



Hình 2. Kết quả huấn luyện mô hình ViT-DFNet

Nhìn chung, mô hình ViT-DFNet của chúng tôi đã thể hiện khả năng tổng quát hóa tốt trên tập dữ liệu kiểm định, với độ chính xác cuối cùng đạt gần 85% và sai số giảm dần rõ rệt về gần 0.3, chứng tỏ cấu trúc Transformer hoạt động hiệu quả trong bài toán này.

3.3.2. Kết quả quá trình đánh giá

Bảng 2 dưới đây trình bày kết quả đánh giá hiệu năng của hai mô hình ResNet50 và ViT-DFNet. Dựa trên các chỉ số độ chính xác Precision (P), độ bao phủ Recall (R) và độ đo Fl-score (F_1), có thể thấy rằng mô hình ViT-DFNet của chúng tôi đạt hiệu quả cao hơn so với ResNet50 trên cả ba tiêu chí. Cụ thể, Precision của ViT-DFNet đạt 0.680, cao hơn 0.665 của ResNet50, cho thấy mô hình đề xuất giảm được số lượng phát hiện sai (false positive). Đồng thời, Recall tăng từ 0.830 lên 0.844, phản ánh khả năng phát hiện chính xác các video giả mạo cao hơn. Nhờ vậy, điểm số F_1 cũng được cải thiện từ 0.739 lên 0.753. Kết quả này chứng minh rằng mô hình ViT-DFNet, với kiến trúc dựa trên Transformer, không chỉ nhận dạng video Deepfake chính xác hơn mà còn duy trì độ tổng quát tốt hơn so với mô hình CNN truyền thống ResNet50.

Bảng 2. Kết quả đánh giá trên tập dữ liệu OpenForensics

Mô hình	TP	FP	TN	FN	P	R	F_1
ResNet50	4 494	2 267	3 325	919	0.665	0.830	0.739
ViT-DFNet	4 568	2 150	3 342	845	**0.680**	**0.844**	**0.753**

Tiếp theo, Bảng 3 trình bày kết quả đánh giá chéo của hai mô hình ResNet50 và ViT-DFNet trên tập dữ liệu 140K-DB, nhằm kiểm tra khả năng tổng quát hóa của mô hình khi áp dụng vào một nguồn dữ liệu hoàn toàn khác so với tập huấn luyện ban đầu. Kết quả cho thấy cả hai mô hình đều có sự suy giảm rõ rệt về các chỉ số Precision (P), Recall (R) và F_1-

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

score (F_1) so với kết quả trước đó trên tập OpenForensics. Cụ thể, mô hình ResNet50 chỉ đạt P = 0.577, R = 0.095, F_1 = 0.163, trong khi mô hình ViT-DFNet dù có hiệu năng tốt hơn nhưng cũng chỉ đạt P = 0.556, R = 0.230, F_1 = 0.326.

Bảng 3. Kết quả đánh giá chéo tập dữ liệu 140K-DB

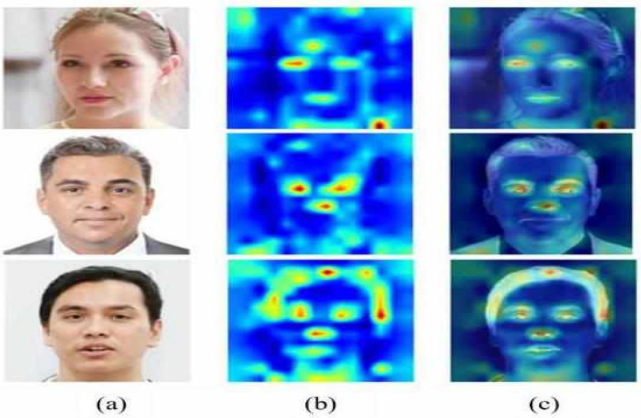
Mô hình	TP	FP	TN	FN	P	R	F_1
ResNet50	954	699	9 301	9 046	0.577	0.095	0.163
ViT-DFNet	2 300	1 836	8 164	7 700	**0.556**	**0.230**	**0.326**

Nguyên nhân của sự suy giảm này có thể được lý giải bởi đặc trưng dữ liệu giữa hai tập huấn luyện và đánh giá có sự khác biệt đáng kể, bao gồm độ phân giải video, định dạng nén, và các thuật toán tạo video Deepfake. Ngoài ra, mô hình ViT-DFNet được tối ưu hóa trên các mẫu giả mạo có đặc điểm khuôn mặt và ánh sáng tương đồng trong OpenForensics, nên khi áp dụng lên 140K-DB - nơi chứa nhiều dạng biến thể mới và kỹ thuật giả mạo phức tạp hơn - hiệu năng bị ảnh hưởng. Mặc dù vậy, ViT-DFNet vẫn duy trì chỉ số F_1 cao gấp đôi ResNet50, cho thấy kiến trúc Transformer của nó có khả năng thích ứng và học đặc trưng tổng quát tốt hơn so với mô hình CNN truyền thống.

Nhìn chung, kết quả thực nghiệm cho thấy mô hình ViT-DFNet đạt hiệu năng cao hơn hơn ResNet50 trên cả hai tập dữ liệu đánh giá. Trên OpenForensics, ViT-DFNet đạt độ chính xác và khả năng tổng quát cao với F_1 = 0.753, trong khi ResNet50 chỉ đạt 0.739. Khi đánh giá chéo trên 140K-DB, cả hai mô hình đều suy giảm hiệu năng do khác biệt về đặc trưng dữ liệu, song ViT-DFNet vẫn duy trì F_1 = 0.326, cao gấp đôi ResNet50. Kết quả này khẳng định khả năng tổng quát và tính ổn định của kiến trúc Transformer trong phát hiện video Deepfake.

3.4. Phân tích khả năng giải thích của mô hình

Trong các mô hình học sâu, quá trình suy luận và ra quyết định thường thiếu tính giải thích, được ví như một "hộp đen", khiến việc phân tích và diễn giải cơ chế hoạt động của mô hình trở nên khó khăn. Để khắc phục vấn đề này, chúng tôi tích hợp thêm kỹ thuật phân tích bản đồ nhiệt nhằm trực quan hóa các vùng mà mô hình tập trung khi đưa ra dự đoán.



Hình 3. Minh họa bản đồ nhiệt của quá trình dự đoán, ảnh đầu vào (a), bản đồ nhiệt (b), ảnh xếp chồng của (a) và (b)

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

Kết quả Hình 3 cho thấy mô hình ViT-DFNet tập trung chủ yếu vào các vùng đặc trưng trên khuôn mặt như mắt, mũi và miệng - những điểm chứa nhiều thông tin quan trọng để phân biệt khuôn mặt thật và giả. Các vùng này thường xuất hiện rõ rệt trên bản đồ nhiệt với màu đỏ hoặc vàng, thể hiện cường độ chú ý cao. Điều này chứng minh rằng mô hình đã học được cách khai thác những chi tiết tinh vi về cấu trúc khuôn mặt. Sự tập trung đúng vào các điểm đặc trưng này phản ánh khả năng học đặc trưng không gian hiệu quả của mô hình ViT-DFNet trong nhiệm vụ nhận dạng khuôn mặt phức tạp.

4. KẾT LUẬN

Mô hình ViT-DFNet mà chúng tôi đề xuất đã chứng minh hiệu quả trong việc phát hiện sự hoán đổi khuôn mặt trong các video được tạo bởi công nghệ Deepfake. Kết quả của mô hình ViT-DFNet cũng cho thấy hiệu suất cao hơn so với các mô hình CNN truyền thống như ResNet50. Đặc biệt, ViT-DFNet có khả năng phát hiện chính xác những sai lệch tinh vi trong khuôn mặt, vốn thường bị bỏ sót bởi mạng CNN. Phân tích diễn giải mô hình cho thấy ViT-DFNet tập trung vào các vùng nhạy cảm như mắt, miệng và đường viền khuôn mặt - nơi thường xuất hiện các hiện tượng biến dạng do quá trình hoán đổi và pha trộn khuôn mặt trong video giả mạo. Kết quả này không chỉ khẳng định ưu thế của kiến trúc Transformer trong bài toán nhận diện nội dung giả mạo, mà còn cung cấp góc nhìn giải thích sâu hơn về cách mô hình hiểu và phân biệt các biến đổi nhân tạo trong dữ liệu video. Hướng nghiên cứu tiếp theo tập trung cải thiện hiệu suất ViT-DFNet bằng cách tối ưu hóa cơ chế chú ý, áp dụng học đa nhiệm và phát triển phiên bản mô hình nhẹ phục vụ phát hiện Deepfake thời gian thực.

TÀI LIỆU THAM KHẢO

[1] Nguyễn Thu Huyền (2023), *Tìm hiểu về một số phương pháp phát hiện ra Deepfake trong Deep learning,* Tạp chí Thiết bị giáo dục, tập 299, kỳ 2, trang 57-59.

[2] Yan, Z., Zhang, Y., Yuan, X., Lyu, S., & Wu, B. D. (2023), *A comprehensive benchmark of deepfake detection,* arXiv preprint arXiv:2307.01426.

[3] Tuan, L. M., Manh, P. T., Linh, D. T. T. (2023), *Deepfake detection based on deep learning,* TNU Journal of Science and Technology, 228(15): 88 - 95.

[4] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C. C. (2020), *The deepfake detection challenge (dfdc) dataset*, arXiv preprint arXiv:2006.07397.

[5] Altuncu, E., Franqueira, V., & Li, S. (2022), *Deepfake: Definitions, performance metrics and standards, datasets and benchmarks, and a meta-review*, arXiv preprint arXiv:2208.10913.

[6] Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., ... & Tao, D. (2024), *Deepfake generation and detection: A benchmark and survey*, arXiv preprint arXiv:2403.17881.

[7] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., ... Nguyen, C. M. (2022), *Deep learning for deepfakes creation and detection: A survey,* Computer Vision and Image Understanding, 223, 103525.

[8] Li, Y., Lyu, S. (2018), *Exposing deepfake videos by detecting face warping artifacts*, arXiv preprint arXiv:1811.00656.

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

[9] Afchar, D., Nozick, V., Yamagishi, J., Echizen, I. (2018), *Mesonet: a compact facial video forgery detection network*, IEEE international workshop on information forensics and security (WIFS) (pp. 1-7).

[10] Kwon, P., You, J., Nam, G., Park, S., & Chae, G. (2021). *Kodf: A large-scale korean deepfake detection dataset,* IEEE/CVF International Conference on computer vision (pp. 10744-10753).

[11] Ni, Y., Meng, D., Yu, C., Quan, C., Ren, D., & Zhao, Y. (2022). *Core: Consistent representation learning for face forgery detection* . IEEE/CVF conference on computer vision and pattern recognition (pp. 12-21).

[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929.

[13] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & NieBner, M. (2019). *Faceforensics++: Learning to detect manipulated facial images.* EEE/CVF international conference on computer vision (pp. 1-11).

[14] X., "140k real and fake faces," 2020. [Online]. Available: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

[15] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). *Celeb-DF: A large-scale challenging dataset for deepfake forensics.* IEEE/CVF conference on computer vision and pattern recognition (pp. 3207-3216).

[16] Le, T. N., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2021). *Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild*, International conference on computer vision - ICCV (pp. 10117-10127).

A TRANSFORMER-BASED APPROACH FOR VIDEO DEEPFAKE DETECTION

Le Van Hao, Tran Doan Minh, Trinh Thi Hop, Le Dieu Linh

ABSTRACT

The rapid advancement of Deepfake technology has made it possible to generate videos with realistic facial manipulations, raising serious concerns about the authenticity of digital content. This work proposes ViT-DFNet, a Transformer-based deep learning model designed for the automatic detection of Deepfake videos. Experimental results indicate that the model achieves more accuracy and more generalization performance when compared with conventional CNN architectures, such as ResNet50. The model effectively identifies fine-grained facial inconsistencies and blending artifacts arising from face-swapping operations. In addition, interpretability analyses are applied to clarify the model's reasoning process, highlighting key facial regions associated with forgery detection.

Keywords: *Deepfake detection, deep learning, image classification.*

Ngày nộp bài: 14/10/2025 ; Ngày gửi phản biện: 23/10/2025; Ngày duyệt đăng: 28/02/2026