Đánh giá và cải tiến hệ thống nhận diện gian lận thi cử theo cử chỉ khuôn mặt bằng kĩ thuật tách nền sử dụng MediaPipe

Huynh Quang Tien
huynhqtienvtag@gmail.com

Tieu Anh Tho
tieuanhtho@gmail.com

Nguyen Quoc Hung
nguyenquochung399@gmail.com

***Abstract*—Đảm bảo tính trung thực trong các kỳ thi trực tuyến đang trở thành một thách thức lớn khi giáo dục từ xa ngày càng phổ biến. Các phương pháp giám sát truyền thống không còn hiệu quả trong môi trường trực tuyến, làm gia tăng nguy cơ gian lận như nhờ trợ giúp từ bên ngoài, thi hộ hoặc sử dụng tài liệu trái phép. Để giải quyết vấn đề này, các hệ thống giám sát tự động dựa trên trí tuệ nhân tạo (AI) đã được phát triển nhằm phát hiện các hành vi bất thường và xác định gian lận trong thi cử. Tuy nhiên, việc đạt được độ chính xác cao trong phát hiện gian lận vẫn là một thách thức lớn do sự thay đổi của môi trường như điều kiện ánh sáng, phông nền và góc quay camera khác nhau. Một trong những trở ngại chính khi huấn luyện các mô hình học máy cho giám sát thi trực tuyến là thiếu tập dữ liệu đa dạng và chất lượng cao. Nếu không có các kỹ thuật tăng cường dữ liệu phù hợp và tối ưu hóa mô hình, hiệu suất của các hệ thống này có thể bị hạn chế. Nghiên cứu này nhằm nâng cao độ chính xác và độ tin cậy của các mô hình học máy trong việc nhận diện hành vi gian lận trong thi cử trực tuyến. Chúng tôi đề xuất một phương pháp tiền xử lý dữ liệu tiên tiến bằng cách tách nền và thay đổi nền, giúp giảm sự phụ thuộc vào yếu tố môi trường và tập trung vào hành vi của thí sinh. Đồng thời, chúng tôi đánh giá hiệu suất của các kiến trúc học sâu, bao gồm Convolutional Neuron Network - CNN và Vision Transformer - ViT, nhằm tìm ra mô hình tối ưu nhất. Kết quả thực nghiệm cho thấy phương pháp xử lý dữ liệu đề xuất giúp cải thiện đáng kể độ chính xác trong phát hiện gian lận. Tuy nhiên, vẫn tồn tại một số thách thức, chẳng hạn như thời gian huấn luyện lâu và tiêu tốn tài nguyên trên máy cá nhân, cũng như lỗi tách nền của MediaPipe, đôi khi không thể tách chính xác khuôn mặt. Bằng cách giải quyết vấn đề thay đổi môi trường trong nhận diện gian lận thi cử trực tuyến, nghiên cứu này cung cấp những hiểu biết quan trọng về tiềm năng của các mô hình học sâu trong việc phát triển hệ thống giám sát hiệu quả hơn.**

***Index Terms*—Online Exam Monitoring, Fraud Detection, Artificial Intelligence, Machine Learning, Deep Learning, Vision Transformer (ViT), Convolutional Neural Networks (CNN), Background Removal, MediaPipe, Proctoring Systems**

I. Giới thiệu

Với sự phát triển nhanh chóng của giáo dục trực tuyến và hình thức kiểm tra từ xa, việc đảm bảo tính trung thực trong các kỳ thi online đang trở thành một thách thức lớn. Các phương pháp giám sát truyền thống không còn phù hợp trong môi trường trực tuyến, làm gia tăng nguy cơ gian lận như nhờ trợ giúp từ bên ngoài, thi hộ, hoặc sử dụng tài liệu trái phép. Để giải quyết vấn đề này, các hệ thống giám sát tự động dựa trên trí tuệ nhân tạo (AI) và học máy (ML) đã ra đời nhằm phát hiện các hành vi bất thường và xác định các trường hợp gian lận trong thi cử.

Mặc dù các hệ thống giám sát dựa trên AI đã đạt được nhiều tiến bộ, nhưng việc phát hiện gian lận một cách chính xác vẫn là một thách thức lớn. Nhiều hệ thống hiện tại có tỷ lệ cảnh báo sai cao do sự thay đổi của môi trường, như điều kiện ánh sáng, phông nền, hoặc góc quay camera khác nhau. Một trong những khó khăn chính trong việc huấn luyện các mô hình ML mạnh mẽ để phát hiện gian lận trong thi cử trực tuyến là thiếu tập dữ liệu đa dạng và chất lượng cao. Nếu không có quá trình tăng cường dữ liệu phù hợp và tối ưu hóa mô hình, các hệ thống giám sát này có thể kém hiệu quả trong thực tế.

Mục tiêu chính của nghiên cứu này là nâng cao độ chính xác và tính tin cậy của các mô hình học máy trong việc nhận diện hành vi gian lận trong các kỳ thi trực tuyến. Chúng tôi đề xuất một kỹ thuật tiền xử lý dữ liệu tiên tiến bằng cách tách nền và thay đổi nền, giúp giảm sự phụ thuộc của mô hình vào yếu tố môi trường và tập trung vào hành vi của thí sinh. Đồng thời, chúng tôi tiến hành đánh giá hiệu suất của các kiến trúc học sâu khác nhau, bao gồm Convolutional Neuron Network - CNN và Vision Transformer - ViT [1], nhằm xác định mô hình tối ưu nhất cho bài toán này.

Bài báo này được tổ chức như sau:

- Phần 2: Trình bày tổng quan về các phương pháp giám sát gian lận thi cử trực tuyến hiện nay.
- Phần 3: Mô tả chi tiết về kỹ thuật tiền xử lý dữ liệu, kiến trúc của các mô hình CNN và ViT, cũng như quy trình huấn luyện mô hình.
- Phần 4: Kết quả so sánh hiệu suất các mô hình và phân tích ảnh hưởng của việc tăng cường dữ liệu.
- Phần 5: Thảo luận về kết quả và hướng nghiên cứu tương lai.

Bằng cách giải quyết các thách thức liên quan đến sự thay đổi môi trường trong nhận diện gian lận thi cử trực tuyến, nghiên cứu này mang lại những hiểu biết quan trọng về tiềm năng của các mô hình học sâu trong việc phát triển hệ thống giám sát hiệu quả hơn.

II. Nghiên cứu liên quan

Bài nghiên cứu này được tiến hành dựa trên phương pháp và dataset từ bài [2]. Ở bài đó, các tác giả đã giới thiệu một hệ thống tên là ExamEdu để nhận diện hành vi gian lận thông qua cử chỉ và cử động đầu bằng mô hình ResNet50 có huấn luyện chuyển tiếp và fine-tuning.

Ngoài ra có nhiều bài nghiên cứu với phương pháp nhận diện tương tự. Chúng ta có thể thấy ở các bài [3], [4] và [5]. Ở bài [6], các tác giả đã giới thiệu mô hình CHEESE để

nhận diện kết hợp tư thế, thông tin khung nền với cử động của mắt, đầu và khuôn mặt. Một bài khác [7], khi các tác giả sử dụng mô hình CNN 3D cùng với OpenCV để nhận diện gian lận trên thời gian thực. Với bài [8], các tác giả đã xây dựng một hệ thống sử dụng một camera CCTV để quan sát khuôn mặt, mắt và thiết bị của học sinh và nhận diện trên thời gian thực qua phương pháp tiếp cận phân cụm - Clustering-based approach. Các tác giả của bài [9] cũng đã giới thiệu một khung phần mềm tên là CLERF để có thể dự đoán chính xác bất kỳ cử động đầu.

Nhìn chung thì việc nhận diện gian lận thi cử trực tuyến là việc nhận diện thông qua dữ liệu về cử động đầu, khuôn mặt và hướng mắt của thí sinh. Chúng tôi dựa trên thông tin đó để xây dựng phương pháp cho hệ thống của chúng tôi để cải thiện hiệu suất nhận diện.

III. Phương pháp

A. Giới thiệu Dataset

Dataset được sử dụng có tất cả 8657 ảnh, được chia thành 3 thư mục "train", "test" và "val". Thư mục "train" được dùng để huấn luyện, trong khi thư mục "test" dùng để kiểm tra độ chính xác trong quá trình huấn luyện. Thư mục "val" sử dụng những hình ảnh không nằm trong quá trình huấn luyện để xác nhận độ chính xác của mô hình sau khi huấn luyện.

Thư mục "train" bao gồm 2 thư mục con là "Cheating" bao gồm 3712 ảnh thí sinh đang gian lận và "Not_Cheating" bao gồm 2688 ảnh thí sinh không gian lận.

Thư mục "test" có cấu trúc giống với thư mục "train", thư mục con "Cheating" bao gồm 736 ảnh, thư mục con "Not_Cheating" bao gồm 394 ảnh.

Thư mục "val", thư mục con "Cheating" bao gồm 734 ảnh, thư mục con "Not_Cheating" bao gồm 393 ảnh.

Fig. 1: Ví dụ về Ảnh đang gian lận

Fig. 2: Ví dụ về Ảnh không gian lận

Ở bài nghiên cứu này, chúng tôi sẽ gọi dataset này là Dataset *Gốc*

B. Tách nền & đổi nền Dataset

Trong việc nhận diện hành vi bất thường trong thi cử, yếu tố môi trường là một trong những trở ngại lớn nhất. Thí sinh có thể thực hiện việc thi ở nhiều không gian khác nhau như nhà riêng, thư viện, quán cà phê,… Điều này làm tăng độ nhiễu của dữ liệu đầu vào, gây khó khăn cho mô hình trong việc nhận diện chính xác hành vi gian lận. Để nâng cao độ chính xác của hệ thống, cần một Dataset có tính đa dạng cao nhưng vẫn đảm bảo sự nhất quán trong việc nhận diện hành vi. Vì vậy, chúng tôi đã áp dụng phương pháp tách nền và đổi nền trên Dataset hiện có nhằm giảm sự nhạy cảm của mô hình đối với yếu tố môi trường xung quanh thí sinh.

a) *Tách nền:*

Chúng tôi sử dụng thư viện MediaPipe [10], một công cụ tiên tiến trong xử lý hình ảnh và nhận diện đối tượng, để xác định chính xác khuôn mặt và cơ thể của thí sinh trong hình ảnh. Quá trình tách nền giúp loại bỏ các yếu tố gây nhiễu không liên quan đến hành vi gian lận, đồng thời chuẩn hóa dữ liệu đầu vào.

Quy trình tách nền bao gồm các bước sau:

- Xác định vùng chứa thí sinh bằng mô hình phân đoạn ảnh từ MediaPipe.
- Tách phần nền khỏi đối tượng chính (thí sinh) bằng thuật toán phân vùng ảnh.
- Xuất ảnh với nền trong suốt để dễ dàng thay đổi nền mới mà không làm mất đi thông tin quan trọng.

Kết quả của quá trình này là một tập hợp hình ảnh chỉ chứa thí sinh với nền trong suốt, giúp mô hình học tập trung vào đặc điểm của thí sinh mà không bị ảnh hưởng bởi môi trường xung quanh.

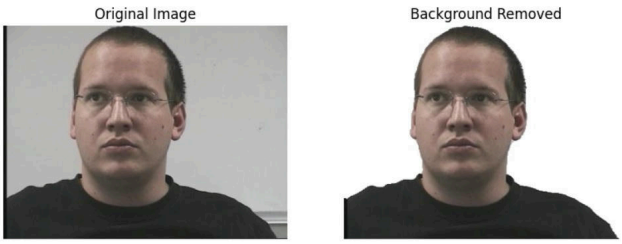

Fig. 3: Ví dụ về Tách nền ảnh

b) *Đổi nền:*

Sau khi tách nền, chúng tôi tiến hành đổi nền để tạo ra bối cảnh mới cho thí sinh. Việc này giúp làm phong phú Dataset và tăng khả năng tổng quát hóa của mô hình.

Quy trình đổi nền bao gồm các bước sau:

- Chúng tôi sử dụng thư viện OpenCV để thực hiện quá trình ghép nền mới với các bước sau:
- Chọn một hình nền đa dạng có độ phức tạp cao, chẳng hạn như thư viện sách với nhiều màu sắc khác nhau.
- Ghép thí sinh vào nền mới bằng cách sử dụng kỹ thuật pha trộn ảnh để đảm bảo tính tự nhiên.
- Điều chỉnh kích thước ảnh sao cho nhất quán với ảnh gốc nhằm duy trì tính đồng nhất của Dataset.

Việc đổi nền giúp mô hình không bị phụ thuộc vào một loại môi trường cố định, từ đó nâng cao khả năng phát hiện hành vi bất thường trong nhiều điều kiện khác nhau.

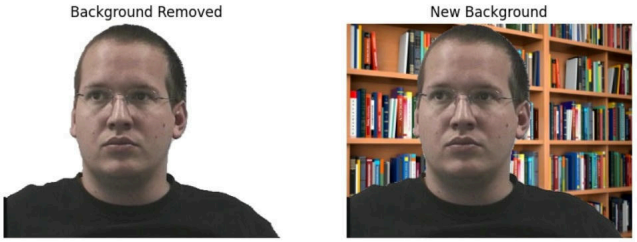

Fig. 4: Ví dụ về Đổi nền ảnh

c) *Lợi ích của phương pháp:*

Phương pháp làm giàu Dataset bằng tách nền và đổi nền mang lại nhiều lợi ích quan trọng:

- Tăng tính đa dạng dữ liệu: Dataset sau khi mở rộng sẽ chứa nhiều bối cảnh khác nhau, giúp mô hình học được cách nhận diện hành vi gian lận bất kể môi trường xung quanh.
- Cải thiện khả năng khái quát hóa: Mô hình giảm sự phụ thuộc vào các yếu tố môi trường, tập trung nhiều hơn vào đặc điểm hành vi của thí sinh.
- Giảm chi phí thu thập dữ liệu thực tế: Thay vì ghi lại dữ liệu từ nhiều môi trường khác nhau, chúng tôi có thể tái sử dụng hình ảnh gốc với các nền khác nhau, tiết kiệm đáng kể thời gian và công sức.
- Tăng độ chính xác của mô hình: Bằng cách loại bỏ nhiễu và cung cấp dữ liệu đa dạng hơn, mô hình học máy có thể phân biệt chính xác hơn giữa hành vi bình thường và gian lận.

d) *Kết quả:*

Việc áp dụng phương pháp tách nền và đổi nền đã mang lại những cải thiện đáng kể trong quá trình xây dựng Dataset phục vụ cho nhận diện hành vi gian lận thi cử. Cụ thể:

- Dataset đã được mở rộng với số lượng ảnh tăng gấp 3 lần.
- Hình ảnh sau xử lý có tính đa dạng cao hơn nhưng vẫn giữ được thông tin quan trọng của thí sinh.
- Các thử nghiệm ban đầu cho thấy mô hình hoạt động ổn định hơn và ít bị ảnh hưởng bởi sự thay đổi môi trường.

Phương pháp này mở ra hướng đi mới trong việc xây dựng Dataset hiệu quả, giảm chi phí và tối ưu hóa nguồn dữ liệu đầu vào cho các hệ thống nhận diện gian lận thi cử.

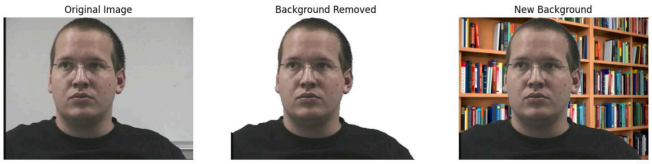

Fig. 5: Kết quả ảnh sau xử lý

Chúng ta sẽ gọi dataset này là Dataset *Tách nền*

C. Huấn luyện mô hình CNN

Để tìm ra mô hình tối ưu nhất cho mục đích của chúng tôi, chúng tôi đã tiến hành thử nghiệm trên bốn mô hình học sâu:

- ResNet50V2 [11]: Kiến trúc ResNet (Residual Network) giúp giảm vấn đề mất mát gradient trong mạng sâu, cải thiện độ chính xác.
- Xception [12]: Sử dụng Depthwise Separable Convolutions, giúp giảm số lượng tham số và cải thiện hiệu suất trong việc trích xuất đặc trưng.
- VGG19 [13]: Có cấu trúc đơn giản, dễ hiểu và hoạt động tốt trên nhiều loại dữ liệu hình ảnh, đặc biệt khi sử dụng với học chuyển giao.
- MobileNetV2 [14]: Nhẹ, nhanh, và được thiết kế tối ưu cho các thiết bị di động hoặc khi tài nguyên tính toán bị hạn chế.

Các mô hình này được lựa chọn do hiệu suất nổi bật của chúng trong các nhiệm vụ phân loại hình ảnh, cũng như sự khác biệt trong kiến trúc, giúp chúng tôi đánh giá tác động của từng mô hình đến kết quả cuối cùng.

a) *Cấu hình huấn luyện:*

Tất cả các mô hình được cấu hình với cùng các siêu tham số để đảm bảo tính nhất quán trong quá trình huấn luyện:

- Số epoch = 50
- Batch size = 16
- Learning rate = 1e-5

Các mô hình được huấn luyện trên cùng một tập con của tập dữ liệu huấn luyện và được đánh giá trên một tập con hình ảnh kiểm tra độc lập.

b) *Học chuyển giao:*

Để tối ưu hóa hiệu suất và giảm thời gian huấn luyện, chúng tôi sử dụng trọng số của các mô hình đã được huấn luyện trước. Cụ thể, các mô hình này đã được huấn luyện trên tập dữ liệu ImageNet, bao gồm 1,28 triệu hình ảnh thuộc 1000 lớp.

Bằng cách sử dụng học chuyển giao, chúng tôi tận dụng các trọng số đã được huấn luyện sẵn thay vì khởi tạo từ đầu,

giúp cải thiện độ chính xác trong khi vẫn sử dụng một tập dữ liệu nhỏ hơn.

Sau khi tải trọng số của mô hình gốc, chúng tôi thay thế lớp fully-connected cuối cùng và lớp đầu ra bằng một lớp fully-connected mới và một lớp đầu ra với hàm kích hoạt softmax để phân loại dữ liệu thành hai lớp: gian lận và không gian lận.

c) *Quy trình huấn luyện:*

Quá trình huấn luyện được chia thành hai giai đoạn chính, được gọi là Giai đoạn A và Giai đoạn B.

- Giai đoạn A
  - Đóng băng các trọng số của mạng nơ-ron gốc và chỉ huấn luyện các lớp fully-connected mới.
  - Mục tiêu là để các lớp mới học được các đặc trưng có ý nghĩa từ dữ liệu đầu vào mà không làm thay đổi các trọng số của mạng gốc.
- Giai đoạn B
  - Mở khóa quá trình học cho toàn bộ mô hình, bao gồm cả các lớp gốc của mạng nơ-ron ban đầu.
  - Huấn luyện toàn bộ mô hình trên tập dữ liệu để tối ưu hóa trọng số của tất cả các lớp.

d) *Kết quả:* Sau khi hoàn tất quá trình huấn luyện và đánh giá trên dataset có tách nền, chúng tôi thu được kết quả như Table I

TABLE I: Kết quả huấn luyện mô hình CNN trên dataset đã tách nền

Mô hình	Độ chính xác
ResNet50V2	0.9782
Xception	0.9826
VGG19	0.9838
MobileNetV2	0.9676

Từ kết quả trên, chúng tôi nhận thấy mô hình VGG19 đạt độ chính xác cao nhất (98.38%), tiếp theo là Xception (98.26%) và ResNet50V2 (97.82%). Mô hình MobileNetV2 mặc dù có độ chính xác thấp hơn (96.76%) nhưng vẫn có ưu thế về tốc độ và hiệu suất tính toán.

Sau đó, chúng tôi cũng tiến hành huấn luyện trên dataset gốc (không tách nền) và thu được kết quả như Table II

TABLE II: Kết quả huấn luyện mô hình CNN trên dataset đã tách nền

Mô hình	Độ chính xác
ResNet50V2	0.9734
Xception	0.9725
VGG19	0.8876
MobileNetV2	0.9778

Từ bảng kết quả trên, chúng tôi nhận thấy mô hình MobileNetV2 đạt độ chính xác cao nhất (97.78%), tiếp theo là ResNet50V2 (97.34%) và Xception (97.25%). Mô hình VGG19 có độ chính xác thấp nhất (88.76%) trên tập dữ liệu gốc.

So sánh hai bảng kết quả, có thể thấy rằng việc tách nền giúp tăng đáng kể độ chính xác của hầu hết các mô hình, đặc biệt là VGG19, khi độ chính xác tăng từ 88.76% lên 98.38%. Trong khi đó, các mô hình như ResNet50V2 và Xception có sự cải thiện nhẹ, cho thấy chúng vẫn hoạt động ổn định trên cả hai loại dữ liệu.

D. Huấn luyện mô hình Visual Transformer

Trong nghiên cứu này, chúng tôi cũng thử nghiệm trên mô hình Vision Transformer - ViT với tập dữ liệu đã có nhằm để đánh giá mô hình với các mô hình CNN ở phần trước.

ViT là một mô hình dựa trên cơ chế tự chú ý (self-attention), giúp mô hình có thể học được mối quan hệ giữa các vùng khác nhau trong hình ảnh, thay vì chỉ dựa vào các bộ lọc cục bộ như CNN.

Ở bài nghiên cứu này, do giới hạn về phần cứng nên chúng tôi sử dụng phiên bản gọn nhẹ hơn của mô hình ViT được gọi là SimpleViT [15].

a) *Cấu hình phần cứng và môi trường huấn luyện:*

Quá trình huấn luyện được thực hiện trên một máy laptop cá nhân chạy hệ điều hành Arch Linux, với GPU NVIDIA GTX 1650. Việc tối ưu hóa và tăng tốc huấn luyện được thực hiện bằng cách sử dụng CUDA để tận dụng khả năng xử lý song song của GPU.

Quá trình huấn luyện và tinh chỉnh mô hình được thực hiện bằng Pytorch và Pytorch Lightning, hai thư viện mạnh mẽ hỗ trợ việc xây dựng và huấn luyện mô hình học sâu một cách linh hoạt và hiệu quả.

b) *Quá trình huấn luyện:*

Chúng tôi thực hiện huấn luyện mô hình theo hai giai đoạn:

- Huấn luyện trên dataset gốc
  - Mô hình ViT được huấn luyện từ đầu trên dataset gốc.
  - Sử dụng tối ưu hóa Adam với learning rate 0.00025 và batch size 22.
  - Mô hình sau khi huấn luyện ở giai đoạn này được gọi là ViT *Gốc*.
- Huấn luyện tiếp tục với dataset đã tách nền
  - Mô hình tiếp tục được huấn luyện (fine-tune) trên tập dữ liệu đã chỉnh sửa nhằm giúp cải thiện khả năng nhận diện gian lận trong môi trường có biến đổi dữ liệu.
  - Learning rate và batch size vẫn giữ nguyên.
  - Mô hình sau khi huấn luyện ở giai đoạn này được gọi là ViT *Tách nền*.

c) *Kết quả:*

Sau khi huấn luyện mô hình, chúng tôi được kết quả như Table III

TABLE III: Độ chính xác sau khi huấn luyện mô hình ViT trên các Dataset

Mô hình	Gốc	Tách nền
ViT *Gốc*	96.5%	90.5%
ViT *Tách nền*	95.6%	95.2%

Các kết quả này cho thấy việc huấn luyện tiếp tục trên tập dữ liệu được chỉnh sửa giúp mô hình thích ứng tốt hơn với các trường hợp gian lận đã được tinh chỉnh, đồng thời vẫn duy trì hiệu suất cao trên tập dữ liệu gốc.

IV. Kết quả

Các thí nghiệm được thực hiện trên tập dữ liệu bao gồm cả ảnh gốc và ảnh đã qua xử lý (tách nền và thay đổi nền). Chúng tôi huấn luyện và đánh giá các mô hình ResNet50V2, Xception, VGG19, MobileNetV2 và ViT bằng cách sử dụng chỉ số đánh giá theo độ chính xác (accuracy) trên tập dữ liệu thử nghiệm. Table IV cho thấy kết quả cuối cùng sau khi huấn luyện các mô hình cho cả tập dữ liệu gốc và đã xử lý.

TABLE IV: Độ chính xác của các mô hình sau huấn luyện trên các Dataset

Mô hình	Gốc	Tách nền
ResNet50V2	0.9734	0.9782
Xception	0.9725	0.9826
VGG19	0.8876	**0.9838**
MobleNetV2	**0.9778**	0.9676
ViT	0.9562	0.9520

Kết quả thực nghiệm chỉ ra rằng phương pháp tách nền và thay đổi nền có tác động tích cực đến hiệu suất nhận diện gian lận. Việc loại bỏ nhiễu môi trường giúp mô hình tập trung hơn vào hành vi của thí sinh, giảm thiểu tỷ lệ cảnh báo sai.

Những kết quả này nhấn mạnh tầm quan trọng của việc lựa chọn mô hình phù hợp kết hợp với kỹ thuật xử lý dữ liệu tiên tiến để nâng cao khả năng phát hiện gian lận trong thi cử trực tuyến.

V. Thảo luận

Mặc dù nghiên cứu đã chứng minh rằng phương pháp tách nền và thay đổi nền có thể cải thiện hiệu suất mô hình, nhưng vẫn tồn tại một số hạn chế.

Thứ nhất, việc huấn luyện mô hình ViT yêu cầu lượng tài nguyên tính toán lớn, khiến thời gian huấn luyện kéo dài đáng kể khi chạy trên máy cá nhân. Điều này làm giảm khả năng mở rộng của mô hình trong môi trường thực tế.

Thứ hai, việc sử dụng MediaPipe để tách nền mặc dù mang lại lợi ích trong việc giảm nhiễu môi trường, nhưng không phải lúc nào cũng đảm bảo chất lượng ảnh sau khi xử lý. Một số trường hợp, công cụ này không thể tách chính xác khuôn mặt, gây ra lỗi trong dữ liệu huấn luyện và có thể làm giảm độ chính xác của mô hình.

Trong tương lai, chúng tôi đề xuất cải thiện phương pháp tách nền bằng cách sử dụng các mô hình phân đoạn ảnh tiên tiến hơn như U-Net hoặc DeepLabV3+. Đồng thời, việc triển khai mô hình trên phần cứng mạnh hơn hoặc sử dụng dịch vụ đám mây sẽ giúp rút ngắn thời gian huấn luyện và cải thiện hiệu suất tổng thể.

References

[1] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." [Online]. Available: https://arxiv.org/abs/2010.11929

[2] H. H. Luong, T. T. Khanh, M. D. Ngoc, M. H. Kha, K. T. Duy, and T. T. Anh, "Detecting Exams Fraud Using Transfer Learning and Fine-Tuning for ResNet50," in *Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications*, T. K. Dang, J. Küng, and T. M. Chung, Eds., Singapore: Springer Nature Singapore, 2022, pp. 747–754.

[3] A. Singh and S. Das, "A Cheating Detection System in Online Examinations Based on the Analysis of Eye-Gaze and Head-Pose," EAI, 2022. doi: 10.4108/eai.16-4-2022.2318165.

[4] I. N. Yulita, F. A. Hariz, I. Suryana, and A. S. Prabuwono, "Educational Innovation Faced with COVID-19: Deep Learning for Online Exam Cheating Detection," *Education Sciences*, vol. 13, no. 2, 2023, doi: 10.3390/educsci13020194.

[5] C. S. Indi, V. Pritham, V. Acharya, and K. Prakasha, "Detection of Malpractice in E-exams by Head Pose and Gaze Estimation," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 8, pp. pp.47–60, Apr. 2021, doi: 10.3991/ijet.v16i08.15995.

[6] Y. Liu, J. Ren, J. Xu, X. Bai, R. Kaur, and F. Xia, "Multiple Instance Learning for Cheating Detection and Localization in Online Examinations," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 4, pp. 1315–1326, Aug. 2024, doi: 10.1109/tcds.2024.3349705.

[7] S. El Kohli, Y. Jannaj, M. Maanan, and H. Rhinane, "DEEP LEARNING: NEW APPROACH FOR DETECTING SCHOLAR EXAMS FRAUD," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 103–107, Jan. 2022, doi: 10.5194/isprs-archives-xlvi-4-w3-2021-103-2022.

[8] S. Z. Ong, T. Connie, and M. K. O. Goh, "Cheating Detection for Online Examination Using Clustering Based Approach," *JOIV : International Journal on Informatics Visualization*, vol. 7, no. 3–2, p. 2075, Nov. 2023, doi: 10.30630/joiv.7.3-2.2327.

[9] T.-R. Wei, H. Liu, H.-C. Hu, X. Wu, Y. Fang, and H.-T. Wu, "CLERF: Contrastive LEaRning for Full Range Head Pose Estimation." [Online]. Available: https://arxiv.org/abs/2412.02066

[10] C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines." [Online]. Available: https://arxiv.org/abs/1906.08172

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks." [Online]. Available: https://arxiv.org/abs/1603.05027

[12] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions." [Online]. Available: https://arxiv.org/abs/1610.02357

[13] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks." [Online]. Available: https://arxiv.org/abs/1801.04381

[15] L. Beyer, X. Zhai, and A. Kolesnikov, "Better plain ViT baselines for ImageNet-1k." [Online]. Available: https://arxiv.org/abs/2205.01580