

LIGHTWEIGHT 3D CNN WITH CHANNEL-SPATIAL ATTENTION FOR CHEATING BEHAVIOR RECOGNITION IN EXAM SURVEILLANCE VIDEOS

Dinh Duc Hoang¹, Dinh Thi Thanh Uyen^{2*}

¹TNU - University of Information and Communication Technology, ²TNU – University of Agriculture and Forestry

ARTICLE INFO	ABSTRACT
Received: 17/5/2026 Revised: 29/6/2026 Published: 30/6/2026	Automatic detection of cheating behaviors in examinations is essential for maintaining academic integrity, particularly in large-scale and online assessments. This study proposes a lightweight human action recognition model that is suitable for deployment on edge devices with limited computational resources. The proposed method extends the MobileNet architecture into a three-dimensional convolutional neural network by employing three-dimensional depthwise separable convolutions to efficiently extract spatial and temporal features from video data. In addition, a convolutional block attention module is incorporated to enhance the model's ability to focus on salient motion regions while suppressing irrelevant background information. Experimental results on an examination surveillance dataset containing six categories of cheating behaviors demonstrate that the proposed model requires only 3.2 million parameters and 1.8 billion floating-point operations for each 16-frame video clip, substantially reducing computational cost compared with the reference model. The model achieves an accuracy of 93.7% and a weighted F1-score of 0.932. These results confirm that the proposed model provides an effective balance between recognition accuracy and computational efficiency, making it well suited for real-time examination surveillance systems deployed on edge devices.
KEYWORDS	
Cheating detection	
Lightweight 3D CNN	
Exam monitoring	
Action recognition	
Attention mechanism	

MẠNG CNN 3D NHẸ VỚI CƠ CHẾ CHÚ Ý THEO KÊNH VÀ KHÔNG GIAN ĐỂ NHẬN DIỆN HÀNH VI GIAN LẬN TRONG VIDEO GIÁM SÁT KỲ THI

Dinh Đức Hoàng¹, Dinh Thị Thanh Uyên^{2*}

¹Trường Đại học Công nghệ Thông tin và Truyền thông - ĐH Thái Nguyên,

²Trường Đại học Nông Lâm - ĐH Thái Nguyên

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 17/5/2026 Ngày hoàn thiện: 29/6/2026 Ngày đăng: 30/6/2026	Phát hiện tự động các hành vi gian lận trong thi cử là yêu cầu quan trọng nhằm bảo đảm tính liêm chính học thuật, đặc biệt trong các kỳ thi trực tuyến và quy mô lớn. Nghiên cứu này đề xuất một mô hình nhận dạng hành động nhẹ, phù hợp triển khai trên các thiết bị biên có tài nguyên tính toán hạn chế. Phương pháp nghiên cứu mở rộng kiến trúc MobileNet sang mạng nơ-ron tích chập ba chiều sử dụng tích chập phân tách theo chiều sâu để trích xuất đặc trưng không gian và thời gian từ video, đồng thời tích hợp mô-đun chú ý khối tích chập nhằm tăng cường khả năng tập trung vào các vùng chuyển động quan trọng và giảm ảnh hưởng của nền. Kết quả thực nghiệm trên bộ dữ liệu giám sát kỳ thi gồm sáu loại hành vi gian lận cho thấy mô hình chỉ sử dụng 3,2 triệu tham số và yêu cầu 1,8 tỷ phép toán dấu phẩy động cho mỗi đoạn video gồm 16 khung hình, giảm đáng kể chi phí tính toán so với mô hình tham chiếu, đồng thời đạt độ chính xác 93,7% và điểm F1 có trọng số đạt 0,932. Kết quả nghiên cứu khẳng định mô hình đề xuất đạt sự cân bằng giữa độ chính xác và hiệu quả tính toán, phù hợp cho các hệ thống giám sát thi cử thời gian thực trên thiết bị biên.
TỪ KHÓA	
Nhận diện hành vi gian lận	
3D CNN nhẹ	
Giám sát thi cử	
Nhận dạng hành động	
Cơ chế chú ý	

DOI: <https://doi.org/10.34238/tnu-jst.15789>

* Corresponding author. Email: dinhthithanhuyen@tuaf.edu.vn

1. Giới thiệu

Gian lận học thuật trong các kỳ thi đặt ra một thách thức dai dẳng đối với các cơ sở giáo dục trên toàn thế giới. Sự mở rộng nhanh chóng của việc kiểm tra trực tuyến và từ xa, được thúc đẩy bởi đại dịch COVID-19, đã làm trầm trọng thêm mối lo ngại về mức độ phổ biến của gian lận [1] [2]. Giám sát thủ công truyền thống, mặc dù hiệu quả trong môi trường quy mô nhỏ, ngày càng trở nên không thực tế khi quy mô lớp học tăng lên và việc giám sát chuyển sang môi trường không được giám sát. Các nghiên cứu gần đây đã ghi nhận tỷ lệ gian lận vượt quá 30% trong các bài đánh giá trực tuyến [3], nhấn mạnh nhu cầu cấp thiết về các giải pháp giám sát tự động. Nhận dạng hành vi gian lận tự động bằng video cung cấp một giải pháp thay thế có thể mở rộng, nhưng đòi hỏi các mô hình có khả năng suy luận thời gian thực trên phần cứng có khả năng tính toán hạn chế được triển khai trong các phòng thi.

Nhận dạng hành động video đã tiến bộ đáng kể thông qua các mạng nơ-ron tích chập sâu. CNN hai chiều kết hợp với các chiến lược tổng hợp thời gian như Mạng phân đoạn thời gian [4] và Mô-đun dịch chuyển thời gian [5] mang lại hiệu quả tính toán nhưng thiếu mô hình thời gian rõ ràng của chuyển động chi tiết. Ngoài ra, các hướng tiếp cận dựa trên phát hiện đối tượng theo từng khung hình (ví dụ như các hệ thống tương tự YOLO) thường yêu cầu pipeline nhiều bước (phát hiện – theo dõi – suy luận hành vi), do đó khó nắm bắt trực tiếp các đặc trưng động học liên tục của hành vi. Mạng CNN ba chiều, bao gồm C3D [6] và I3D [7], kết hợp chiều thời gian trực tiếp vào các bộ lọc tích chập, đạt được độ chính xác vượt trội với chi phí tính toán đáng kể. Tích chập (2+1)D trong [8] được đề xuất để giảm chi phí tính toán so với tích chập 3D từ C3D và I3D. Mạng SlowFast [9] đã cải thiện hơn nữa độ chính xác thông qua xử lý đường dẫn kép nhưng vẫn yêu cầu hơn 34 triệu tham số. Kích thước mô hình và số lượng FLOP cao khiến các kiến trúc này không phù hợp để triển khai trên các thiết bị có tài nguyên hạn chế như máy tính cá nhân, NVIDIA Jetson hoặc Raspberry Pi, điển hình cho các hệ thống giám sát trong lớp học.

Các kiến trúc nhẹ đã xuất hiện để thu hẹp khoảng cách giữa hiệu quả và độ chính xác. MobileNet [10] đã giới thiệu các phép tích chập tách biệt theo chiều sâu để giảm đáng kể số tham số trong khi vẫn duy trì độ chính xác cạnh tranh. Các phần mở rộng gần đây cho video, bao gồm TSM-MobileNetV3 [11] và MNv3-MFAE [12], đã chứng minh tính khả thi của các thiết kế thân thiện với thiết bị di động cho nhận dạng hành động. Tuy nhiên, các phương pháp chuyển kênh cung cấp trường tiếp nhận thời gian hạn chế, và bản chất tinh tế, chi tiết của các hành vi gian lận—những cử chỉ tay ngắn gọn, những cái quay đầu lén lút và những thao tác vật thể nhỏ—đòi hỏi các mô hình có khả năng mô hình hóa không gian–thời gian rõ ràng.

Các cơ chế chú ý như Squeeze-and-Excitation (SE) [13] và Convolutional Block Attention Module (CBAM) [14] đã được chứng minh là hiệu quả trong việc tăng cường khả năng phân biệt đặc trưng với chi phí tối thiểu [15], [16]. Tuy nhiên, trong phần lớn các nghiên cứu hiện có, CBAM chủ yếu được tích hợp vào các kiến trúc CNN 2D (ví dụ như các biến thể ResNet-CBAM), do đó chỉ khai thác hiệu quả thông tin không gian mà chưa tận dụng đầy đủ chiều thời gian của dữ liệu video. Bên cạnh đó, việc kết hợp cơ chế chú ý với các kiến trúc 3D nhẹ nhằm đồng thời đảm bảo khả năng biểu diễn không gian–thời gian và hiệu quả tính toán vẫn còn là một hướng nghiên cứu chưa được khai thác đầy đủ, đặc biệt trong bài toán phát hiện hành vi gian lận trong môi trường giám sát thi cử.

Trong bài báo này, chúng tôi đề xuất MCNet-CBAM, một CNN 3D nhẹ giải quyết những thách thức trên bằng cách kết hợp hiệu quả mô hình hóa không gian–thời gian với thiết kế tối ưu cho thiết bị biên. Khác với các phương pháp 2D kết hợp chú ý hoặc các pipeline phát hiện đối tượng theo từng khung hình, mô hình đề xuất học trực tiếp biểu diễn hành vi từ chuỗi video thông qua tích chập 3D và cơ chế chú ý mở rộng theo không gian–thời gian. Những đóng góp của chúng tôi như sau:

- Chúng tôi thiết kế một khối tích chập tách biệt theo chiều sâu 3D mở rộng triết lý MobileNet [10] cho video, đạt được số lượng tham số chỉ 3,2 triệu.

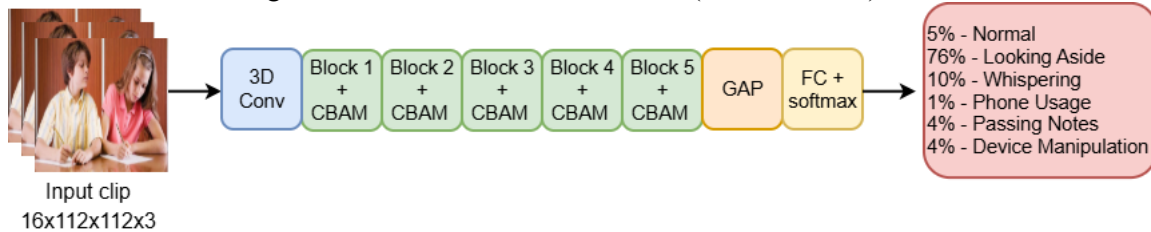
- Chúng tôi tích hợp CBAM [14] vào kiến trúc 3D, cung cấp sự chú ý theo kênh và không gian tuần tự chỉ với 0,08 triệu tham số bổ sung.
- Chúng tôi xây dựng bộ dữ liệu ExaCheat-6 và đánh giá so với C3D [6] và I3D [7], đạt được độ chính xác 93,7% với mức giảm tham số gấp 12,4 lần so với C3D.
- Chúng tôi tiến hành các đánh giá toàn diện về loại bỏ và hiệu quả bao gồm kích thước mô hình, FLOPs và suy luận trên phần cứng biên.

2. Phương pháp

Trong phần này, chúng tôi trình bày chi tiết về kiến trúc mô hình MCNet-CBAM, tập trung vào việc áp dụng các khối tích chập 3D nhẹ kết hợp cơ chế chú ý nhằm đạt được hiệu suất tối ưu trên phần cứng bị hạn chế.

2.1. Tổng quan về Kiến trúc MCNet-CBAM

MCNet-CBAM bao gồm 5 khối tích chập tách biệt chiều sâu 3D (3D depthwise separable convolution blocks). Để tối ưu hoá khả năng tập trung vào các đặc trưng quan trọng, mô-đun CBAM được tích hợp ngay sau 4 khối đầu tiên. Đầu vào của mạng là một clip gồm $T=16$ khung hình, kích thước khung hình $H=W=112$, với số kênh $C=3$ (ảnh màu RGB).



Hình 1. Tổng quan mô hình MCNet-CBAM.

Trong đó, 3D Conv là phép tích chập 3D, GAP là tầng Global Average Pooling

Kiến trúc mạng sử dụng các khối tích chập tách biệt chiều sâu 3D (3D Depthwise Separable Convolution) để giảm thiểu số lượng tham số trong khi vẫn giữ được khả năng trích xuất đặc trưng không gian - thời gian từ video. Mỗi khối bao gồm một lớp tích chập chiều sâu (Depthwise Convolution) với kernel $3 \times 3 \times 3$, theo sau là một lớp tích chập điểm (Pointwise Convolution) với kernel $1 \times 1 \times 1$. Cơ chế CBAM được tích hợp ngay sau lớp Pointwise Convolution để tinh chỉnh các đặc trưng. Bảng 1 mô tả kiến trúc chi tiết của mô hình được đề xuất MCNet-CBAM.

Bảng 1. Kiến trúc chi tiết của MCNet-CBAM

Khối	Tham số	Số lần lặp	Kích thước đầu ra
Input Block (3D Conv)	$5 \times 7 \times 7$, stride 2, 32 channels	1	$16 \times 56 \times 56 \times 32$
Block 1 (DW + PW + CBAM)	$3 \times 3 \times 3$, stride 1, 64 channels	1	$16 \times 56 \times 56 \times 64$
Block 2 (DW + PW + CBAM)	$3 \times 3 \times 3$, stride 2, 128 channels	1	$8 \times 28 \times 28 \times 128$
Block 3 (DW + PW + CBAM)	$3 \times 3 \times 3$, stride 2, 256 channels	1	$4 \times 14 \times 14 \times 256$
Block 4 (DW + PW + CBAM)	$3 \times 3 \times 3$, stride 2, 512 channels	1	$2 \times 7 \times 7 \times 512$
Block 5 (DW + PW)	$3 \times 3 \times 3$, stride 1, 1024 channels	1	$2 \times 7 \times 7 \times 1024$
Global Average Pooling	-	1	$1 \times 1 \times 1 \times 1024$
Fully Connected	1024 $\rightarrow$ 6 classes	1	6

Như được mô tả trong Hình 1, toàn bộ pipeline xử lý của MCNet-CBAM như sau: Input Block $\rightarrow$ Block 1 (DW+PW+CBAM) $\rightarrow$ Block 2 (DW+PW+CBAM) $\rightarrow$ Block 3 (DW+PW+CBAM) $\rightarrow$ Block 4 (DW+PW+CBAM) $\rightarrow$ Block 5 (DW+PW) $\rightarrow$ GAP $\rightarrow$ FC(1024 $\rightarrow$ 6) $\rightarrow$ Softmax $\rightarrow$ Dự đoán.

Xét khối Block 1 (các khối khác tương tự), giả sử đầu vào là tensor 5 chiều: $X \in R^{T \times H \times W \times C}$. Trong đó: T: chiều sâu (Depth/ Time), H: chiều cao (Height), W: chiều rộng (Width), C: số kênh (Channels). Đối với DW, mỗi kênh được tích chập độc lập với một kernel riêng với kernel có kích thước như trong công thức (1).

$$K(c) \in R^{3 \times 3 \times 3}, c = 1, 2, \dots, C \quad (1)$$

Đầu ra:

$$Z \in R^{T' \times H' \times W' \times C} \quad (2)$$

Công thức cho DW được tính như sau:

$$Z(t, h, w, c) = \sum_{i=0}^2 \sum_{j=0}^2 \sum_{k=0}^2 K(c)(i, j, k) \cdot X(t + i, h + j, w + k, c) \quad (3)$$

Đối với DW, để giảm thiểu số lượng phép tính toán mà vẫn giữ được khả năng học hỏi đặc trưng từ đầu vào, các khối DW được thiết kế để không có sự tương tác giữa các kênh và do đó, số lượng kernel sẽ bằng chính số kênh C .

Đối với PW, lớp này thực hiện tích chập $1 \times 1 \times 1$ để kết hợp thông tin giữa các kênh. Gọi Kernel có kích thước như sau:

$$P \in R^{1 \times 1 \times 1 \times C \times C'} \quad (4)$$

Trong đó: C' là số kênh đầu ra và Đầu ra cuối có kích thước như sau:

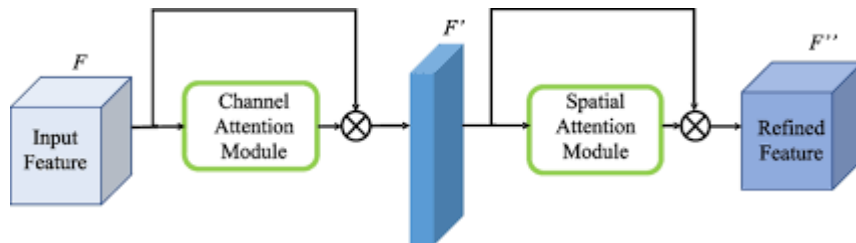
$$Y \in R^{T' \times H' \times W' \times C'} \quad (5)$$

Kết quả đầu ra Y được tính theo công thức:

$$Y(t, h, w, c') = \sum_{c=1}^C P(c, c') \cdot Z(t, h, w, c) \quad (6)$$

Khối 3D Depthwise Separable Convolution phân tách phép tích chập 3D truyền thống thành hai phép toán tuần tự: (i) tích chập chiều sâu (depthwise convolution) với kernel $3 \times 3 \times 3$ áp dụng độc lập trên từng kênh để trích xuất đặc trưng không gian–thời gian cục bộ; và (ii) tích chập điểm (pointwise convolution) với kernel $1 \times 1 \times 1$ nhằm học sự kết hợp tuyến tính giữa các kênh.

2.2. Cơ chế chú ý CBAM (Convolutional Block Attention Module)



Hình 2. Cấu trúc của khối chú ý CBAM

Mô đun CBAM (Hình 2) hoạt động theo trình tự hai bước cốt lõi: Attention theo kênh (Channel Attention) và Attention theo không thời gian (Spatial Attention). Cơ chế này xử lý bản đồ đặc trưng F đầu vào thông qua các phép toán tính toán chi tiết như sau:

- Channel Attention: Nhằm khai thác thông tin biểu diễn giữa các kênh của đặc trưng. Phép toán sử dụng Average Pooling và Max Pooling trên không gian để tạo ra hai vector đặc trưng riêng biệt, sau đó đưa qua một mạng nơ-ron đa tầng (MLP) chung. Đầu ra được kết hợp lại để tạo ra bản đồ chú ý kênh M_c . Lớp Channel Attention giúp mạng lưới tập trung vào các bản đồ đặc trưng mang tính quyết định cao nhất đối với hành vi gian lận.

- Spatial Attention: Hoạt động trên các đặc trưng đã được tinh chỉnh bởi Channel Attention. Nó áp dụng Average Pooling và Max Pooling trên chiều kênh, tiếp theo là phép tích chập (ví dụ: kernel $7 \times 7 \times 7$ cho 3D) để xác định các vị trí không gian quan trọng. Lớp này định hướng sự chú ý của mô hình vào các vùng không gian cụ thể như chuyển động của tay, hướng mắt nhìn thay vì bối cảnh xung quanh.

Để tăng cường khả năng biểu diễn đặc trưng của mô hình 3D CNN nhẹ trong bài toán nhận diện hành vi gian lận, chúng tôi tích hợp mô-đun Convolutional Block Attention Module

(CBAM) vào kiến trúc mạng. CBAM là một cơ chế attention tuần tự, bao gồm hai thành phần chính: Channel Attention và Spatial Attention, giúp mô hình học được “nên chú ý vào cái gì” và “chú ý ở đâu” trong không gian–thời gian. Giả sử đầu vào của mô-đun là tensor đặc trưng $F \in R^{T \times H \times W \times C}$, trong đó T, H, W, C lần lượt là chiều thời gian, chiều cao, chiều rộng và số kênh.

Channel Attention: Mục tiêu của Channel Attention là khai thác mối quan hệ phụ thuộc giữa các kênh đặc trưng, từ đó nhấn mạnh các kênh mang thông tin quan trọng đối với việc phân biệt hành vi. Cụ thể, tensor F được nén theo chiều không gian–thời gian thông qua hai phép pooling toàn cục: Global Average Pooling và Global Max Pooling, thu được hai vector đặc trưng:

$$F_{avg}^c, F_{max}^c \in R^{1 \times 1 \times 1 \times C} \quad (7)$$

Hai vector này sau đó được đưa qua một mạng MLP dùng chung tham số, gồm hai lớp fully connected với tỉ lệ giảm chiều r, nhằm học biểu diễn phi tuyến giữa các kênh. Kết quả từ hai nhánh được cộng phần tử và đưa qua hàm kích hoạt sigmoid để tạo bản đồ attention theo kênh:

$$Mc(F) = \sigma(MLP(F_{avg}^c) + MLP(F_{max}^c)) \quad (8)$$

Bản đồ attention này được sử dụng để tái cân bằng tensor đầu vào:

$$F' = Mc(F) \odot F \quad (9)$$

trong đó $\odot$ là phép nhân từng phần tử có broadcast. Cơ chế này cho phép mô hình ưu tiên các kênh đặc trưng mang tính phân biệt cao, ví dụ như các kênh liên quan đến chuyển động tay, hướng nhìn hoặc tư thế đầu trong bối cảnh giám sát gian lận.

Spatial Attention: Sau khi tinh chỉnh theo kênh, mô-đun Spatial Attention tiếp tục xác định các vùng không gian–thời gian quan trọng trong tensor đặc trưng. Đầu tiên, tensor F' được nén theo chiều kênh bằng hai phép pooling:

$$F_{avg}^s, F_{max}^s \in R^{T \times H \times W \times 1} \quad (10)$$

Hai bản đồ này được nối theo chiều kênh, tạo thành tensor:

$$[F_{avg}^s, F_{max}^s] \quad (11)$$

Sau đó, một lớp tích chập 3D với kernel kích thước $7 \times 7 \times 7$ được áp dụng nhằm học các mẫu phụ thuộc không gian–thời gian cục bộ. Kết quả được đưa qua hàm sigmoid để tạo bản đồ attention không gian:

$$Ms(F') = \sigma(f^{7 \times 7 \times 7}([F_{avg}^s, F_{max}^s])) \quad (12)$$

Cuối cùng, tensor đầu ra được tính như sau:

$$F'' = Ms(F') \odot F' \quad (13)$$

Cơ chế này giúp mô hình tập trung vào các vùng quan trọng như chuyển động bất thường của tay, ánh mắt hoặc tương tác đáng ngờ, đồng thời giảm nhiễu từ các vùng nền không liên quan.

- Việc tích hợp CBAM vào kiến trúc được thúc đẩy bởi các ưu điểm sau:
- Khả năng chọn lọc đặc trưng hiệu quả: Channel Attention giúp xác định các kênh quan trọng nhất đối với từng loại hành vi gian lận.
- Định vị không gian–thời gian chính xác: Spatial Attention tập trung vào các vùng có chuyển động bất thường, loại bỏ ảnh hưởng của nền.
- Chi phí tính toán thấp: CBAM chỉ làm tăng khoảng 0,08 triệu tham số (dưới 3% tổng số tham số), phù hợp với mục tiêu xây dựng mô hình nhẹ.
- Dễ tích hợp: Mô-đun có thể chèn trực tiếp vào các khối tích chập 3D mà không yêu cầu thay đổi cấu trúc backbone, đồng thời hoạt động theo cơ chế tự giám sát.

3. Thử nghiệm

Trong phần này, chúng tôi trình bày về bộ dữ liệu được sử dụng để thử nghiệm và chi tiết triển khai mô hình MCNet-CBAM. Sau đó, chúng tôi cung cấp các kết quả và phân tích so sánh giữa các mô hình. Cuối cùng là các phân tích cắt bỏ nhằm thấy được sự hiệu quả của khối CBAM mang lại.

3.1. Bộ dữ liệu

Tập dữ liệu ExaCheat-6 được xây dựng có hệ thống nhằm phục vụ bài toán nhận diện hành vi gian lận trong môi trường thi cử. Bộ dữ liệu bao gồm 4.800 đoạn video ngắn (clips), được thu thập từ 40 người tham gia, mô phỏng 6 loại hành vi diễn hình trong phòng thi, bao gồm: Normal (hành vi bình thường), Looking Aside (nhìn ngang/dọc), Whispering (thì thầm), Phone Usage (sử dụng điện thoại), Passing Notes (chuyên giấy), và Device Manipulation (tương tác với thiết bị khác). Để thu thập được bộ dữ liệu này, chúng tôi tuân thủ các nguyên tắc đạo đức trong thu thập và xử lý dữ liệu con người. Tất cả 40 người tham gia đều được thông báo rõ ràng về mục đích nghiên cứu và đã cung cấp sự đồng ý tự nguyện trước khi ghi hình. Dữ liệu thu thập được chỉ sử dụng cho mục đích nghiên cứu, được ẩn danh hóa và không chứa thông tin nhận dạng cá nhân. Quy trình thu thập và sử dụng dữ liệu đảm bảo tuân thủ các quy định hiện hành về bảo vệ quyền riêng tư.

Mỗi đoạn video có độ dài 2 giây, được ghi hình ở tốc độ 30 khung hình/giây (FPS). Từ mỗi clip, chúng tôi thực hiện trích mẫu 16 khung hình, sau đó chuẩn hóa về kích thước không gian 112×112 pixel, phù hợp với đầu vào của mô hình 3D CNN. Tập dữ liệu được phân chia theo tỷ lệ 70% cho huấn luyện, 15% cho kiểm thử (validation) và 15% cho đánh giá (test), đảm bảo tính khách quan trong quá trình huấn luyện và đánh giá mô hình. Để đánh giá tác động của độ dài chuỗi video đến hiệu quả nhận diện hành vi gian lận, chúng tôi tiến hành một thí nghiệm bổ sung với các độ dài clip khác nhau, bao gồm 1 giây (8 khung hình), 2 giây (16 khung hình) và 4 giây (32 khung hình), trong đó tốc độ lấy mẫu được giữ cố định. Kết quả cho thấy, với độ dài clip ngắn (1 giây), mô hình có xu hướng bỏ sót các hành vi có tính liên tục hoặc diễn ra chậm theo thời gian, chẳng hạn như hành vi “nhìn sang bên” hoặc “trao đổi nhỏ”. Điều này là do thông tin ngữ cảnh theo thời gian chưa được tích lũy đầy đủ, dẫn đến giảm độ chính xác tổng thể. Ngược lại, khi tăng độ dài clip lên 4 giây, mô hình có khả năng nắm bắt tốt hơn các đặc trưng động học dài hạn, tuy nhiên chi phí tính toán tăng đáng kể, làm giảm tốc độ suy luận (FPS), ảnh hưởng đến khả năng triển khai thời gian thực trên các thiết bị biên. Đáng chú ý, với độ dài 2 giây (16 khung hình), mô hình đạt được sự cân bằng tốt giữa độ chính xác và hiệu năng. Độ dài này đủ để bao quát các chuyển động quan trọng trong hầu hết các hành vi gian lận phổ biến, đồng thời vẫn đảm bảo tốc độ xử lý cao.

Về phân bố dữ liệu, mỗi lớp hành vi bao gồm 800 clips, được thu thập trong nhiều điều kiện khác nhau nhằm tăng tính đa dạng, bao gồm sự thay đổi về góc quay, ánh sáng và bố cục phòng thi. Các đối tượng tham gia được chia thành 4 nhóm, mỗi nhóm gồm 10 người, với các kịch bản phòng thi mô phỏng khác nhau. Dữ liệu video được ghi bằng camera độ phân giải 1080p đặt tại góc phòng, nhằm đảm bảo bao quát toàn bộ không gian quan sát.

Quá trình gán nhãn được thực hiện bởi ba chuyên gia độc lập, và mỗi mẫu được chấp nhận khi có ít nhất 2/3 số phiếu đồng thuận. Độ tin cậy giữa các chuyên gia được đánh giá thông qua hệ số Cohen's Kappa đạt 0,92, cho thấy mức độ nhất quán cao trong quá trình chú thích dữ liệu.

Để cải thiện khả năng tổng quát hóa của mô hình, chúng tôi áp dụng một số kỹ thuật tăng cường dữ liệu (data augmentation), bao gồm:

- (1) Lật ngang ngẫu nhiên với xác suất 0,5;
- (2) Điều chỉnh độ sáng trong khoảng $\pm 10\%$;
- (3) Dịch chuyển thời gian trong phạm vi ± 2 khung hình;
- (4) Cắt ngẫu nhiên (random crop) vùng 112×112 từ ảnh gốc kích thước 128×128 .

Các phép biến đổi này giúp làm phong phú không gian dữ liệu đầu vào, tương đương với việc tạo ra hơn 10.000 mẫu hiệu dụng trong mỗi epoch, từ đó góp phần nâng cao độ ổn định và hiệu quả học của mô hình.

3.2. Triển khai

Mô hình được triển khai bằng framework PyTorch (phiên bản 2.0) và huấn luyện trên GPU NVIDIA RTX A4000, tận dụng các thư viện tăng tốc tính toán chuyên dụng cho học sâu nhằm

nâng cao hiệu quả xử lý. Đối với bài toán phân loại đa lớp, hàm mất mát Cross-Entropy. Quá trình tối ưu hóa được thực hiện thông qua thuật toán Adam với các siêu tham số $\beta_1 = 0,9$, $\beta_2 = 0,999$. Tốc độ học ban đầu được thiết lập là 1×10^{-3} và được điều chỉnh theo chiến lược cosine annealing, giảm dần về 1×10^{-6} , giúp cải thiện khả năng hội tụ và tránh rơi vào cực tiểu cục bộ. Đồng thời, hệ số weight decay được đặt ở mức 1×10^{-4} nhằm giảm thiểu hiện tượng overfitting. Kích thước batch được lựa chọn là 32, đảm bảo cân bằng giữa hiệu năng tính toán và độ ổn định của quá trình huấn luyện.

Các mô hình lightweight đề xuất được huấn luyện từ đầu (from scratch) trong 100 epochs. Trong khi đó, các mô hình tham chiếu như C3D và I3D được khởi tạo từ trọng số đã huấn luyện trước trên tập dữ liệu Sports-1M, sau đó được tinh chỉnh (fine-tuning) trong 50 epochs để phù hợp với bài toán cụ thể.

Về chi phí tính toán, thời gian huấn luyện trung bình được ghi nhận như sau: mô hình đề xuất MCNet-CBAM mất khoảng 6 giờ cho 100 epochs (huấn luyện từ đầu), trong khi C3D và I3D lần lượt yêu cầu khoảng 8 giờ và 12 giờ cho 50 epochs fine-tuning. Ngoài ra, các thí nghiệm triển khai suy luận (inference) trên máy tính với CPU i3 3,0 GHz với độ chính xác số học FP16 cho thấy khả năng ứng dụng thực tiễn của mô hình trong các hệ thống giám sát thời gian thực.

3.3. Các chỉ số đánh giá

Để đánh giá và so sánh hiệu năng của các mô hình, chúng tôi sử dụng tập các chỉ số phổ biến trong bài toán phân loại đa lớp, bao gồm Accuracy, Precision, Recall và F1-score. Bên cạnh các chỉ số đánh giá độ chính xác, chúng tôi cũng xem xét các tiêu chí về hiệu quả tính toán nhằm đánh giá tính khả thi trong triển khai thực tế, bao gồm:

- Số lượng tham số (Params): tổng số tham số có thể huấn luyện của mô hình, phản ánh độ phức tạp về bộ nhớ.
- Số phép toán (FLOPs): được tính trên mỗi đầu vào là một clip gồm 16 khung hình, đại diện cho chi phí tính toán.
- Tốc độ xử lý (FPS – Frames Per Second): được đo trên máy tính thông thường (CPU i3) với độ chính xác số học FP16 và kích thước batch bằng 1, nhằm đánh giá khả năng suy luận thời gian thực.

Các tiêu chí trên cung cấp một đánh giá toàn diện, không chỉ về độ chính xác phân loại mà còn về hiệu năng và khả năng triển khai của mô hình trong các hệ thống giám sát thực tế.

3.4. Phân tích so sánh

Kết quả thực nghiệm trong Bảng 2 cho thấy mô hình đề xuất MCNet-CBAM đạt độ chính xác (Acc) 93,7%, cao hơn 5,1 điểm phần trăm so với kiến trúc 2D CNN tiêu chuẩn dựa trên TSN (ResNet-50). Điều này khẳng định vai trò quan trọng của việc mô hình hóa đặc trưng không gian-thời gian trong bài toán nhận diện hành vi.

Bảng 2. So sánh hiệu năng với các mô hình cơ sở

Mô hình	Acc (%)	Prec (%)	Rec (%)	F1	Params	GFLOPs	FPS
TSN (ResNet-50) [4]	88,6	87,9	88,2	0,880	24,6M	33,0	45
C3D [6]	94,5	94,1	94,3	0,942	39,6M	16,8	18
I3D [7]	94,2	93,8	94,0	0,939	34,0M	43,2	12
TSM-MobileNetV3 [11]	91,8	91,3	91,5	0,914	3,5M	2,1	78
MCNet-CBAM (Ours)	93,7	93,3	93,5	0,932	3,2M	1,8	85

Khi so sánh với các kiến trúc 3D CNN quy mô lớn như C3D và I3D, mô hình MCNet-CBAM chỉ suy giảm nhẹ về độ chính xác (khoảng 0,5–0,8%). Tuy nhiên, đổi lại, mô hình đạt được sự cải thiện đáng kể về hiệu quả tính toán: số lượng tham số giảm tới 12,4 lần so với C3D và chi phí tính toán (FLOPs) giảm khoảng 9,3 lần. Kết quả này cho thấy việc kết hợp 3D depthwise separable convolution với cơ chế attention CBAM mang lại khả năng cân bằng hiệu quả giữa độ chính xác và chi phí tính toán.

So với mô hình nhẹ TSM-MobileNetV3, MCNet-CBAM đạt độ chính xác cao hơn 1,9%, đồng thời vẫn duy trì số lượng tham số thấp hơn. Điều này cho thấy ưu thế của việc khai thác trực tiếp đặc trưng không gian–thời gian bằng tích chập 3D kết hợp attention, thay vì dựa trên các cơ chế xấp xỉ theo thời gian.

Về khả năng triển khai thời gian thực, MCNet-CBAM đạt tốc độ 85 FPS trên máy tính CPU i3 thông thường với FP16, batch = 1, vượt xa ngưỡng yêu cầu 30 FPS cho các hệ thống giám sát trực tuyến. Trong khi đó, các mô hình như C3D (18 FPS) và I3D (12 FPS) không đáp ứng được yêu cầu thời gian thực, còn TSM-MobileNetV3 tuy đạt tốc độ cao (78 FPS) nhưng có độ chính xác thấp hơn.

Những kết quả này cho thấy MCNet-CBAM đạt được sự cân bằng tối ưu giữa độ chính xác, độ phức tạp tính toán và tốc độ suy luận, qua đó phù hợp cho các ứng dụng giám sát gian lận trong môi trường thực tế có hạn chế về tài nguyên phần cứng.

Bảng 3. Chi tiết ma trận nhầm lẫn

Actual \ Predicted	Normal	Looking Aside	Whispering	Phone Usage	Passing Notes	Device Manip.
Normal	118	1	1	0	0	0
Looking Aside	4	106	7	0	2	1
Whispering	1	8	102	0	5	4
Phone Usage	0	0	0	119	1	0
Passing Notes	0	1	2	1	115	1
Device Manip.	0	1	1	1	2	115

Để hiểu rõ hơn về hiệu năng của mô hình, chúng tôi tiến hành phân tích ma trận nhầm lẫn (ở Bảng 3) trên tập kiểm tra nhằm xác định các loại hành vi dễ bị dự đoán sai. Kết quả cho thấy, mặc dù mô hình đạt độ chính xác tổng thể cao, vẫn tồn tại một số cặp hành vi có xu hướng bị nhầm lẫn do sự tương đồng về đặc trưng không gian–thời gian.

Cụ thể, hành vi “nhìn sang bên” (Looking Aside) đôi khi bị nhầm với các hành vi có chuyển động đầu nhẹ hoặc không rõ ràng, chẳng hạn như “gãi đầu”, do cả hai đều liên quan đến sự thay đổi vị trí vùng đầu trong thời gian ngắn. Tương tự, hành vi “thì thầm” (Whispering) có thể bị nhầm lẫn với các hành vi tương tác nhẹ khác khi khoảng cách giữa các đối tượng nhỏ hoặc khi chuyển động miệng không đủ rõ ràng trong video có độ phân giải thấp.

Ngoài ra, các hành vi có đặc trưng tĩnh vi và thời gian diễn ra ngắn thường có tỷ lệ nhầm lẫn cao hơn so với các hành vi có chuyển động rõ ràng và kéo dài. Điều này cho thấy mặc dù mô hình đã khai thác tốt thông tin không gian–thời gian, việc phân biệt các hành vi vi mô (micro-actions) vẫn là một thách thức, đặc biệt trong điều kiện dữ liệu nhiễu hoặc thiếu thông tin.

3.5. Phân tích Cắt bỏ

Để đánh giá một cách định lượng đóng góp của từng thành phần trong kiến trúc đề xuất, chúng tôi tiến hành nghiên cứu cắt bỏ trên các biến thể của mô hình. Kết quả trong Bảng 4 cho thấy, khi loại bỏ hoàn toàn cơ chế attention (mô hình Baseline), độ chính xác chỉ đạt 89,5%.

Bảng 4. Kết quả nghiên cứu cắt bỏ

Cấu hình	Acc (%)	Δ Baseline	Params	GFLOPs
Baseline (không attention)	89,5	-	3,12M	1,72
Baseline + SE (chi channel)	92,1	+2,6	3,16M	1,76
Baseline + CBAM spatial-only	91,6	+2,1	3,16M	1,77
Baseline + CBAM (channel+spatial)	93,7	+4,2	3,20M	1,80
Baseline + CBAM (2 lần)	93,8	+4,3	3,28M	1,88

Việc bổ sung riêng lẻ cơ chế chú ý theo kênh thông qua Squeeze-and-Excitation (SE) giúp cải thiện độ chính xác lên 92,1%, tương ứng mức tăng 2,6 điểm phần trăm so với baseline. Tương tự, khi chỉ sử dụng thành phần spatial attention trong CBAM, mô hình đạt 91,6%, tăng 2,1%, cho thấy cả hai cơ chế chú ý đều đóng vai trò độc lập và có ý nghĩa trong việc nâng cao hiệu năng.

Khi kết hợp đồng thời channel attention và spatial attention trong cấu trúc hoàn chỉnh của CBAM, độ chính xác đạt 93,7%, tăng 4,2% so với baseline. Kết quả này chứng minh sự bổ trợ lẫn nhau (complementarity) giữa hai cơ chế attention, giúp mô hình học được cả “đặc trưng quan trọng” (what) và “vị trí quan trọng” (where) trong dữ liệu không gian-thời gian.

Ngoài ra, việc chèn CBAM hai lần liên tiếp chỉ mang lại cải thiện rất nhỏ (+0,1%) trong khi làm tăng số lượng tham số, cho thấy cấu hình một mô-đun CBAM là đủ để đạt hiệu quả tối ưu.

4. Kết luận

Trong nghiên cứu này, chúng tôi đề xuất mô hình MCNet-CBAM, một kiến trúc 3D CNN nhẹ được thiết kế chuyên biệt cho bài toán nhận diện hành vi gian lận trong môi trường thi cử. Mô hình đạt hiệu năng cao với độ chính xác 93,7%, trong khi chỉ sử dụng 3,2 triệu tham số và 1,8 GFLOPs, đồng thời đạt tốc độ suy luận 85 FPS trên máy tính cá nhân thông thường. So với các mô hình 3D CNN truyền thống như C3D, MCNet-CBAM giúp giảm kích thước mô hình tới 12,4 lần nhưng vẫn duy trì độ chính xác cạnh tranh. Những kết quả này cho thấy hiệu quả của việc kết hợp 3D depthwise separable convolution với cơ chế attention CBAM trong việc đạt được sự cân bằng giữa độ chính xác và chi phí tính toán, phù hợp với các hệ thống giám sát thời gian thực có hạn chế về tài nguyên. Trong tương lai, chúng tôi định hướng mở rộng nghiên cứu theo các hướng sau: (1) Mô hình hóa phụ thuộc thời gian dài hạn thông qua việc tích hợp các kiến trúc như Transformer hoặc LSTM nhằm khai thác ngữ cảnh chuỗi tốt hơn; (2) Kết hợp dữ liệu từ nhiều góc nhìn để nâng cao độ bao quát và độ chính xác nhận dạng; (3) Tối ưu hóa triển khai trên thiết bị biên thông qua các kỹ thuật nén mô hình như lượng tử hóa nhằm giảm tiêu thụ năng lượng.

TÀI LIỆU THAM KHẢO/REFERENCES

- [1] P. M. Newton and K. Essex, “How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review,” *J. Academic Ethics*, vol. 22, no. 2, pp. 323-343, 2024.
- [2] T. T. T. Phung, N. K. Duong, and P. L. C. Nguyen, “Recognizing cheating behavior in exams from camera images based on YOLOv8 and ResNet-CBAM models,” *TNU J. Sci. Technol.*, vol. 230, no. 15, pp. 178-185, 2025.
- [3] T. Jannat, H. Ullah, M. Hosen, N. A. Omar, S. Arefin, and J. Hazelton, “Cracking the code: Unveiling e-cheating trends among tertiary education students,” *Tertiary Educ. Manag.*, vol. 31, no. 4, pp. 1-25, 2026.
- [4] L. Wang *et al.*, “Temporal segment networks: Towards good practices for deep action recognition,” in *Proc. ECCV*, 2016, pp. 20-36.
- [5] D. Q. Vu and T. T. T. Phung, “Simultaneous context and motion learning in video prediction,” *Signal Image Video Process.*, vol. 17, no. 8, pp. 3933-3942, 2023.
- [6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in *Proc. ICCV*, 2015, pp. 4489-4497.
- [7] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in *Proc. CVPR*, 2017, pp. 6299-6308.
- [8] T. Phung, V. T. Nguyen, T. H. T. Ma, and D. Q. Vu, “A (2+ 1) d attention convolutional neural network for video prediction,” in *International Conference on Artificial Intelligence and Big Data in Digital Era*, 2021, pp. 395-406.
- [9] C. Feichtenhofer, H. Fan, J. Malik, and K. He, “SlowFast networks for video recognition,” in *Proc. ICCV*, 2019, pp. 6202-6211.
- [10] A. G. Howard *et al.*, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017, doi: 10.48550/arXiv.1704.04861.
- [11] Y. Wang, L. Chen, and R. Zhang, “TSM-MobileNetV3: A novel lightweight network model for video action recognition,” *IEEE Access*, vol. 11, pp. 88123-88135, 2023.
- [12] Z. Tan, W. Li, and J. Chen, “MNv3-MFAE: A lightweight network for video action recognition,” *Electronics*, vol. 14, no. 5, 2025, Art. no. 981.
- [13] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. CVPR*, 2018, pp. 7132-7141.
- [14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. ECCV*, 2018, pp. 3-19.
- [15] D. Q. Vu, N. T. H. Le, and J. C. Wang, “(2+ 1) d distilled shufflenet: A lightweight unsupervised distillation network for human action recognition,” in *26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 3197-3203.
- [16] D. Q. Vu, T. T. H. Nguyen, D. V. Nguyen, Y. Q. Nguyen, T. N. Phung, and T. T. T. Phung, “A Lightweight Attention Model for Face Recognition,” in *International Conference on Advances in Information and Communication Technology*, 2023, pp. 219-227.