TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 22, Số 5 (2025): 801-813
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 22, No. 5 (2025): 801-813
https://doi.org/10.54607/hcmue.js.22.5.4699(2025)

Bài báo nghiên cứu

MÔ HÌNH C-ViDNet HỖ TRỢ PHÁT HIỆN BẠO LỰC TRONG HỌC ĐƯỜNG

Nguyễn Viết Hưng, Tạ Công Phi*, Lê Tấn Lộc, Ngô Quang Khánh, Trần Thanh Nhã
Trường Đại học Sư phạm Thành phố Hồ Chí Minh, Việt Nam
**Tác giả liên hệ: Tạ Công Phi – Email: tacongphi1@gmail.com*
Ngày nhận bài: 20-01-2025; ngày nhận bài sửa: 23-4-2025; ngày duyệt đăng: 24-4-2025

TÓM TẮT

Bạo lực học đường là một vấn đề phức tạp và đáng lo ngại trong hệ thống giáo dục của nhiều quốc gia trên thế giới, trong đó có Việt Nam. Mặc dù đã có nhiều mô hình phát hiện bạo lực tự động được phát triển dựa trên trí tuệ nhân tạo, nhưng việc triển khai thực tế vẫn còn gặp nhiều khó khăn do độ phức tạp và chi phí tính toán lớn. Để khắc phục các hạn chế này, nghiên cứu của chúng tôi đề xuất xây dựng một mô hình C-ViDNet (Campus Violence Detection Network) phát hiện bạo lực học đường tự động với số lượng tham số nhỏ nhằm tăng cường khả năng phát hiện và phản ứng nhanh với các vụ việc bạo lực trong môi trường giáo dục. Đầu tiên, YOLOX được sử dụng để xác định của những người xuất hiện trong khung hình. Tiếp theo, tư thế của những người này được trích xuất bằng HRNet và chuyển đổi thành 3D Heatmap Volumes, giúp giảm nhiễu và loại bỏ các yếu tố nền không cần thiết. Sau đó, một kiến trúc gồm hai luồng được triển khai để học các đặc trưng từ 3D Heatmap Volumes. Trong đó, một luồng tập trung vào đặc trưng không gian của tư thế, trong khi luồng còn lại theo dõi sự thay đổi tư thế của con người giữa các khung hình. Kết quả từ C-ViDNet cho thấy tiềm năng trong việc phát triển mô hình phát hiện bạo lực học đường tự động. Giải pháp này không chỉ giảm bớt sự phụ thuộc vào giám sát thủ công mà còn đảm bảo phát hiện kịp thời các tình huống bạo lực, hỗ trợ nhà trường trong việc xây dựng môi trường an toàn hơn cho học sinh.

Từ khóa: bạo lực; bạo lực học đường; nhận dạng hành vi; thị giác máy tính; xử lí ảnh; Yolo

1. Giới thiệu

Bạo lực nói chung đã được xác định rõ ràng là một vấn đề nghiêm trọng đối với sức khỏe cộng đồng (Rutherford et al., 2007). Không có quốc gia hoặc cộng đồng nào không bị ảnh hưởng bởi bạo lực (World Health Organization, Regional Office for the Eastern Mediterranean, 2024). Bạo lực còn đặc biệt nguy hiểm trong môi trường học đường, nơi học sinh, sinh viên đang trong giai đoạn phát triển quan trọng về cả thể chất lẫn tâm lí. Theo khoản 5 Điều 2 của Nghị định 80/2017/NĐ-CP, bạo lực học đường là hành vi ngược đãi, đánh đập, bạo hành; làm tổn hại đến sức khỏe, thân thể; sỉ nhục, lăng mạ đến danh dự và

Cite this article as: Nguyen, V. H., Ta, C. P., Le, T. L., Ngo, Q. K., & Tran, T. N. (2025). C-ViDNet: amodel for supporting violence detection in schools. *Ho Chi Minh City University of Education Journal of Science, 22*(5), 801-813. https://doi.org/10.54607/hcmue.js.22.5.4699(2025)

nhân phẩm; tẩy chay, cô lập, ruồng rẫy và những hành động gây ảnh hưởng nặng nề tới sức khỏe tinh thần và thể chất của bạn học trong các tổ chức, cơ sở giáo dục (Government, 2017). Trên toàn cầu, bạo lực học đường là một vấn đề nghiêm trọng, với khoảng một nửa số học sinh trong độ tuổi 13-15, tương đương khoảng 150 triệu trẻ em, báo cáo rằng họ đã trải qua bạo lực giữa các bạn đồng trang lứa trong và xung quanh trường học. Đáng lo ngại hơn, khoảng 720 triệu trẻ em trong độ tuổi đi học sống ở các quốc gia mà luật pháp không bảo vệ các em khỏi hình phạt thể xác ở trường (UNICEF, 2021).

Bạo lực học đường gây ra nhiều hậu quả nghiêm trọng, ảnh hưởng sâu sắc đến cả nạn nhân, thủ phạm và môi trường giáo dục. Về mặt tâm lí, nạn nhân thường phải đối mặt với các vấn đề như trầm cảm, lo âu, và sợ hãi, dẫn đến cảm giác cô lập, có nguy cơ cao phát triển các rối loạn tâm thần, thậm chí là tự tử. Đối với thủ phạm, việc tham gia vào các hành vi bạo lực có thể hình thành các hành vi lệch lạc, tạo tiền đề cho các hành động vi phạm pháp luật trong tương lai. Ngoài ra, bạo lực học đường còn tác động tiêu cực đến kết quả học tập. Nạn nhân thường mất tập trung, giảm hứng thú với việc học và có xu hướng bỏ học, trong khi thủ phạm cũng có thể bị gián đoạn quá trình học tập do phải đối mặt với các biện pháp kỉ luật từ nhà trường và pháp luật.

Việc xây dựng một môi trường học đường an toàn, lành mạnh là điều kiện tiên quyết để đảm bảo chất lượng giáo dục. Hiện nay, các mô hình học sâu như mạng nơ-ron tích chập (CNN) và mạng nơ-ron hồi quy (RNN), đã cho thấy hiệu quả cao trong nhiều nhiệm vụ như nhận dạng hành động, giám sát an ninh và phát hiện bất thường trong đám đông. Do đó, có thể ứng dụng học sâu để xây dựng một hệ thống có khả năng phát hiện và cảnh báo nhanh chóng các tình huống bạo lực, góp phần bảo vệ và duy trì môi trường học đường an toàn.

Trong nghiên cứu này, mô hình C-ViDNet được đề xuất để phát hiện bạo lực học đường qua các tác động vật lí giữa các cá nhân (dưới 7 người), dựa trên chuỗi khung hình từ video giám sát. Phương pháp này sử dụng các 3D heatmap volumes, cho phép trích xuất đặc trưng tư thế của người trong không gian và thời gian mà không cần thông tin phông nền. Quy trình này lần lượt phát hiện người thông qua YOLOX, ước lượng tư thế bằng HRNet và chuyển đổi thành 3D Heatmaps Volume cho mỗi tư thế. Sau đó, các 3D heatmap này được xử lí qua một kiến trúc gồm hai luồng, với một luồng trích xuất đặc trưng không gian từ tư thế và một luồng trích xuất đặc trưng thời gian từ sự thay đổi của tư thế. Đặc trưng từ cả hai luồng được kết hợp qua lớp Fusion và tiến hành phân loại để dự đoán xác suất video có chứa bạo lực.

2. Đối tượng và phương pháp nghiên cứu

Phần này tập trung trình bày hai nội dung chính: 1) Đối tượng nghiên cứu: Bao gồm việc phát hiện hành vi bạo lực học đường thông qua dữ liệu từ camera giám sát, cùng với việc nghiên cứu và áp dụng các mô hình học sâu trong nhận diện bạo lực; và 2) Phương pháp nghiên cứu: Trình bày chi tiết quy trình xây dựng mô hình C-ViDNet nhằm hỗ trợ hiệu quả trong việc phát hiện bạo lực trong môi trường học đường.

2.1. Đối tượng nghiên cứu

Các phương pháp truyền thống trong việc phát hiện và ngăn chặn bạo lực chủ yếu dựa vào con người, nhưng chúng tồn tại nhiều hạn chế, chẳng hạn như khó phát hiện các hành vi tinh vi, không thể theo dõi liên tục, và dễ bị chi phối bởi yếu tố chủ quan. Trong khi đó, sự phát triển mạnh mẽ của các hệ thống camera giám sát hiện đại đã trở thành công cụ quan trọng trong việc đảm bảo an ninh, đặc biệt là phát hiện bạo lực (Parui et al., 2023). Tuy nhiên, sự gia tăng nhanh chóng của lượng dữ liệu từ các hệ thống này đã đặt ra thách thức lớn trong việc xây dựng các phương pháp tự động, có khả năng xử lí và phân tích hàng triệu video với độ chính xác cao và tốc độ nhanh chóng (Kumar et al., 2023). Đặc biệt, các hệ thống trường học ngày càng được trang bị camera giám sát để đảm bảo an ninh học đường. Trong bối cảnh này, sự tiến bộ vượt bậc của trí tuệ nhân tạo đã mở ra cơ hội phát triển các giải pháp tự động, hiệu quả hơn để phát hiện và ngăn chặn bạo lực học đường.

Trong vài năm qua, đã có nhiều nghiên cứu tập trung vào bài toán phát hiện bạo lực trong video. Một hướng nghiên cứu đáng chú ý là phát hiện bạo lực dựa trên phân tích âm thanh từ video, cho phép nhận diện hành vi ngay cả khi hình ảnh không rõ ràng (Yildiz et al., 2023; Santos et al., 2021). Đồng thời, các phương pháp phát hiện bạo lực dựa trên hình ảnh đã đạt được những tiến bộ lớn nhờ sự phát triển của thị giác máy tính và xử lí video (Divya et al., 2024; Ghalley et al., 2024).

Các mô hình học sâu như CNN và 3D-CNN đã được sử dụng để phân tích các khung hình và video, nhận diện chính xác các hành vi bạo lực như đánh, đấm, hay xô đẩy dựa trên các đặc trưng trực quan. Ngoài ra, sự kết hợp giữa phân tích âm thanh và hình ảnh đã dẫn đến việc phát triển các mô hình đa phương thức, tận dụng thông tin từ cả hai nguồn dữ liệu (Khan et al., 2024; Wu et al., 2023). Những mô hình này không chỉ tăng cường độ chính xác mà còn cải thiện khả năng phát hiện bạo lực trong các tình huống phức tạp, nơi dữ liệu từ một nguồn đơn lẻ là không đủ.

2.2. Phương pháp phát hiện bạo lực trong học đường

Nghiên cứu này đề xuất C-ViDNet – mô hình có khả năng phát hiện bạo lực học đường qua các tác động vật lí giữa các cá nhân trong một nhóm nhỏ dựa trên chuỗi khung hình từ video giám sát. C-ViDNet sử dụng phương pháp tiền xử lí đầu vào của mô hình PoseC3D (Duan et al., 2022) - 3D heatmap volumes, làm dữ liệu đầu vào cho các mô hình phát hiện bạo lực theo chuỗi khung hình. Phương pháp này giúp mô hình học được sự thay đổi không gian và thời gian thông qua việc sắp xếp tuần tự các tư thế của con người, cũng như giúp giảm bớt đặc trưng đầu vào bằng cách loại bỏ phông nền và chỉ tập trung vào sự biến đổi của tư thế.

- *Xây dựng Limb Pseudo Heatmap*

Đầu tiên chúng tôi tiến hành xây dựng các Limb Pseudo Heatmap (bản đồ nhiệt mô phỏng chi) để biểu diễn trực quan vị trí và phạm vi ảnh hưởng của một đoạn chi cụ thể (như cẳng tay, đùi, cẳng chân) trong một khung hình ảnh. Quy trình này gồm ba bước chính: Phát hiện người (Human Detection), Ước lượng tư thế (Pose Estimation) và Chuyển đổi sang 3D Heatmaps Volume (Khối bản đồ nhiệt 3D).

Hình 1. *Minh họa chi tiết cho 17 Limb Pseudo Heatmap*

Để phát hiện người trong từng khung hình, mô hình YOLOX được sử dụng. Sau khi hoàn thành bước Human Detection, mỗi người xuất hiện trong khung hình sẽ được khoanh vùng bằng một bounding box (hộp giới hạn). Các bounding box này sẽ là đầu vào cho mô hình HRNet, mô hình có khả năng dự đoán chính xác các điểm chính (keypoints) trong không gian cho nhiều người cùng lúc. Sau đó, thông tin các keypoints từ mỗi khung hình được chuyển thành các Heatmap 2D theo chi – gọi là Limb Pseudo Heatmaps, có kích thước $H \times W$, trong đó H và W là chiều dài và chiều rộng tương ứng của mỗi Limb Pseudo Heatmap. Trong trường hợp có hai người trở lên, các chi tương ứng sẽ được biểu diễn trên cùng một Limb Pseudo Heatmap. Ví dụ, nếu trong một khung hình có hai người bắt tay nhau, thì sẽ có một Limb Pseudo Heatmap l_i duy nhất biểu diễn cánh tay của cả hai người.

Với mỗi khung hình, có $k(x_{k,}y_{k,}c_k)$ keypoints, trong đó $x_{k,}y_k$ lần lượt tọa độ x, y trong không gian 2D của khung hình và c_k là độ tin cậy (confidence score) cho biết mức độ chính xác của việc phát hiện keypoint này. Đầu tiên, với mỗi khung hình, chúng tôi tạo K Limb Pseudo Heatmap $L \in [0]^{K\times H\times W}$với K là số lượng limb (chi), Giá trị của Limb Pseudo Heatmap L sẽ được cập nhật theo công thức (1):

$$\mathrm{L_{kij}} = \mathrm{e}^{-\frac{\mathrm{D}\left((\mathrm{i,j}),\mathrm{seg}[\mathrm{a_k},\mathrm{b_k}]\right)^2}{2\times\sigma^2}\times\min\left(\mathrm{c_{a_k}},\mathrm{c_{b_k}}\right)} \quad (1)$$

trong đó, $k \in [0, K]$, $i \in [0, H]$, $j \in [0, W]$, limb thứ k tương ứng với khung xương được tạo bởi hai keypoint a_k, b_k, với D là hàm tính khoảng cách từ điểm (i, j) đến đường nối $seg[a_k, b_k]$ (hay $seg\left[(x_{a_k}, y_{a_k}), (x_{b_k}, y_{b_k})\right]$) và σ (mặc định là 0.6) được sử dụng để kiểm soát phân phối giá trị của các điểm quanh keypoint (phân phối chuẩn). c_{a_k}, c_{b_k} là độ tin cậy của hai keypoint a_k, b_k.

Các Limb Pseudo Heatmap sẽ được điều chỉnh kích thước cho phù hợp và sắp xếp chồng lên nhau theo thời gian T. Quá trình này tạo ra các 3D Heatmap Volume cho từng limb k. Đầu ra cuối cùng sẽ có kích thước tổng là $T \times K \times W \times H$. Hình 1 minh họa chi tiết cho 17 Limb Pseudo Heatmap.

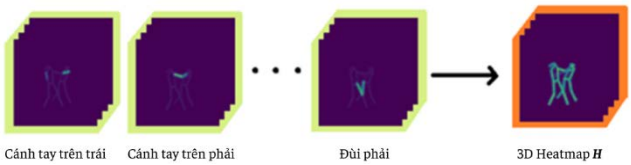
- *Two-Stream Network*

C-ViDNet gồm hai luồng được lấy cảm hứng từ nghiên cứu (Islam et al., 2021) để phát hiện bạo lực trong học đường thông qua dữ liệu video. Đối với luồng đầu tiên, 17 3D

Heatmap Volume có số chiều là $K \times T \times W \times H$ với $K = 17$ sẽ chuyển đổi thành một 3D Heatmap $\boldsymbol{H}$ duy nhất có số chiều là $T \times W \times H$ được tính toán theo công thức (2):

$$\mathbf{H}_{\mathrm{tij}} = \max(\mathrm{L}^{1}_{\mathrm{tij}}, \mathrm{L}^{2}_{\mathrm{tij}}, \dots, \mathrm{L}^{\mathrm{k}}_{\mathrm{tij}}) \quad (2)$$

trong đó, $\boldsymbol{H}_{tij}$ là pixel thứ (i, j) của heatmap thứ t và L^{k}_{tij} là pixel thứ (i, j) của heatmap thứ t của Limb Pseudo Heatmap thứ k trong 3D Heatmap Volume với $t \in [0, T]$, $k \in [1, K]$, $i \in [0, W]$ và $j \in [0, H]$. Đối với luồng này, mục tiêu không phải giảm ảnh hưởng các đặc trưng của ảnh nền, thay vào đó luồng này sẽ tập trung vào việc học tập các đặc trưng về không gian như tư thế, vị trí của người. Hình 2 minh họa chi tiết về 3D Heatmap $\boldsymbol{H}$.



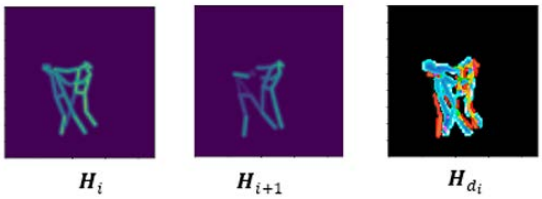
Hình 2. *Minh họa cho 3D Heatmap* ***H***

Luồng thứ hai sẽ tập trung vào sự thay đổi giữa các heatmap liền kề trong $\boldsymbol{H}$ nhằm trích xuất được những thông tin về thời gian như chuyển động, các thay đổi về tư thế người được gọi là Differences Heatmap được biểu diễn như trong công thức (3):

$$\mathrm{H}_{\mathrm{d_i}} = \mathrm{H}_{\mathrm{i+1}} - \mathrm{H}_{\mathrm{i}} \quad (3)$$

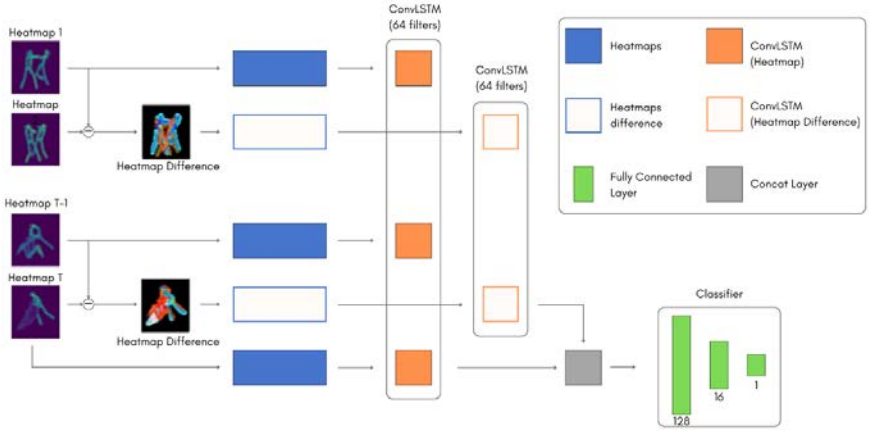
trong đó, H_i là heatmap thứ i từ 3D Heatmap $\boldsymbol{H}$ đầu vào và H_{d_i} là Differences Heatmap biểu diễn sự khác nhau giữa hai heatmap thứ i và $i + 1$, nếu $\boldsymbol{H}$ có t heatmap thì sẽ có $t - 1$ sự khác nhau như thế. Hình 3 minh họa chi tiết về Differences Heatmap H_{d_i}.

Kích thước mỗi Limb Pseudo Heatmap là 224 × 224. Tuy nhiên, để kết hợp thêm đặc trưng về màu sắc nhằm mục đích tăng cường các đặc trưng về sự thay đổi giữa các heatmap, các Limb Pseudo Heatmap được áp dụng colormap viridis (bản đồ màu viridis) và có số chiều là 224 × 224 × 3. Colormap này sử dụng các màu xanh lá cây và xanh dương để biểu diễn dữ liệu thấp và các màu vàng và cam để biểu diễn các giá trị dữ liệu cao hơn trong bản đồ nhiệt (heatmap).



Hình 3. *Minh họa cho Differences Heatmap* H_{d_i}.

Mạng huấn luyện trước MobileNet-v2 sẽ trích xuất hình đặc trưng từ dữ liệu đầu vào của mỗi luồng. MobileNet-V2 sử dụng các lớp depthwise separable convolutions (phép tích chập tách theo chiều sâu) thay vì tích chập thông thường, giúp giảm số lượng tham số và tăng hiệu quả tính toán. Sau đó, các đặc trưng từ mỗi luồng sẽ được đưa qua một lớp ConvLSTM với 64 filter (bộ lọc), kích thước đầu ra 7 × 7 × 64. Tiếp tục qua một lớp Max-Pooling với window size (2,2) để giảm kích thước đặc trưng không gian mà không mất nhiều thông tin quan trọng.



Hình 4. *Kiến trúc mô hình C-ViDNet*

Đặc trưng (feature) được trích xuất từ hai luồng sẽ được tổng hợp ở lớp Concat bằng cách nối (concatenate) hai đặc trưng đầu ra của hai luồng để tạo ra đặc trưng cuối cùng (như trong công thứ 4).

$$\mathrm{F_{fused}} = \mathrm{Concat}(\mathrm{F_h}, \mathrm{F_{diff}}) \quad (4)$$

trong đó, F_h, F_{diff} lần lượt là đặc trưng từ luồng heatmap và luồng sự khác nhau của heatmap tương ứng. F_{fused} là đặc trưng đầu ra của lớp Concat, chứa lần lượt các đặc trưng của F_h, F_{diff} và có kích thước bằng tổng kích thước của F_h và F_{diff}. Việc này giúp tạo nên một bộ đặc trưng hợp nhất phong phú và mạnh mẽ hơn so với việc chỉ xem xét riêng lẻ từng yếu tố. Nhờ đó, mô hình không chỉ đơn thuần biết về tư thế hay chuyển động, mà còn có thể hiểu được cách chúng tương tác và ảnh hưởng lẫn nhau. Các đặc trưng này được chuyển tiếp vào phần phân loại (classifier), bao gồm các Fully Connected Layer (Lớp kết nối đầy đủ - FC) sử dụng hàm kích hoạt LeakyReLU (với hệ số $\alpha = 0.1$) giúp tăng tính phi tuyến tính cho mạng và giảm nguy cơ các neuron không hoạt động. Để giảm nguy cơ quá khớp (overfitting), mỗi lớp được áp dụng Dropout với tỉ lệ 0.3. Lớp cuối cùng là một lớp FC với hàm kích hoạt Sigmoid, đưa ra xác suất dự đoán video đầu vào có chứa nội dung bạo lực hay không. Hàm mất mát Cross-entropy được sử dụng trong suốt quá trình huấn luyện mô hình. Cross-Entropy không chỉ đo lường mức độ chính xác của dự đoán mà còn khuyến khích mô hình cải thiện các dự đoán sai một cách mạnh mẽ và hiệu quả. Điều này tạo ra một hiệu suất huấn luyện rất tốt, đặc biệt trong các trường hợp bài toán phân loại phức tạp như nhận diện bạo lực trong video. Chi tiết về kiến trúc mô hình C-ViDNet được minh họa trong Hình 4.

- *Bộ dữ liệu*

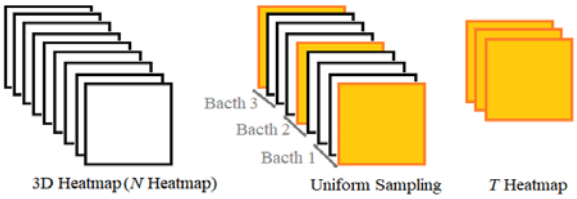
Để huấn luyện và đánh giá hiệu suất của C-ViDNet, chúng tôi sử dụng bốn bộ dữ liệu điểm chuẩn được sử dụng rộng rãi trong các nghiên cứu phát hiện bạo lực, gồm Hockey Fight (Bermejo Nievas et al., 2011), Violence in Movie (VM) (Bermejo Nievas et al., 2011), Real life violence situations (RLVS) (Soliman et al., 2019), Automatic violence detection (AVD) (Bianculli et al., 2020). Chi tiết về các bộ dữ liệu được trình bày như trong Bảng 1. Để việc so sánh với các nghiên cứu trước đó được công bằng và khách quan, chúng tôi không

áp dụng thêm bất kì bước xử lí dữ liệu nào như làm sạch, lọc hay cân bằng lại tỉ lệ giữa các lớp. Bước tiền xử lí được giới hạn ở việc chuẩn hóa kích thước khung hình để đảm bảo tính đồng nhất của dữ liệu đầu vào cho mô hình.

Bảng 1. *Thống kê số lượng video của các bộ dữ liệu*

Bộ dữ liệu	Tập huấn luyện	Tập xác thực	Tập kiểm thử	Tổng
HF	700	200	100	1000
VM	140	40	20	200
RLVS	1400	400	200	2000
AVD	245	70	35	350

Để xác định kích thước đầu vào cho mô hình vì số lượng heatmap trong mỗi 3D Heatmap không đồng đều, cần sử dụng một phương pháp lấy mẫu T heatmap. Do đó, chúng tôi áp dụng phương pháp Uniform Sampling để tăng khả năng nắm bắt thông tin về toàn bộ video. Chiến lược Uniform Sampling tỏ ra hiệu quả hơn trong việc duy trì thông tin chuyển động toàn cục của video, thông qua thực nghiệm cho thấy phương pháp này mang lại lợi thế đáng kể trong nhiệm vụ nhận dạng hành động dựa trên dữ liệu khung xương (Duan et al., 2022). Để lấy mẫu một T heatmap từ 3D Heatmap, phương pháp Uniform Sampling chia 3D Heatmap thành T đoạn có độ dài bằng nhau và lấy mẫu ngẫu nhiên một heatmap từ mỗi đoạn. Như vậy, với phương pháp này, mỗi 3D Heatmap $\boldsymbol{H}$ sẽ trích 32 heatmap để huấn luyện và kiểm thử. Hình 5 minh họa cho kĩ thuật lất mẫu Uniform Sampling được sử dụng trong nghiên cứu này.



Hình 5. *Minh họa cho kĩ thuật lấy mẫu Uniform Sampling*

- *Cài đặt thực nghiệm*

Nghiên cứu này sử dụng ngôn ngữ lập trình Python 3.7.12 cùng thư viện hỗ trợ TensorFlow. Chi tiết về thiết bị huấn luyện và thực nghiệm được trình bày trong Bảng 2. Được truyền cảm hứng từ các nghiên cứu cơ sở trước đây (Kumar et al., 2023; Li et al., 2019; Huszár et al., 2023), chúng tôi tiến hành huấn luyện mô hình trong 50 epochs (kỉ nguyên) với learning rate (tỉ lệ học) là 0.0001 batch size (kích thước lô) là 16 sử dụng optimizer (tối ưu hóa) là Adam. Trong suốt quá trình huấn luyện, mô hình có hiệu suất tốt nhất được lưu lại để tiến hành kiểm tra độ chính xác trên tập dữ liệu kiểm thử.

Bảng 2. *Thiết bị huấn luyện và thực nghiệm*

	Thiết bị huấn luyện	Thiết bị thực nghiệm
Môi trường	Kaggle	Google Colab
Processor	Intel(R) Xeon(R) CPU @ 2.00GHz	Intel(R) Xeon(R) CPU @ 2.00GHz
Memory	13GB RAM	13GB RAM
VGA	Tesla P100 16GB	Tesla T4 15GB

- *Phương pháp đánh giá*

Hiệu suất của mô hình được đánh giá qua hai độ đo là Accuracy và F1 Score. Các số liệu này được tính toán bằng các công thức (6), (8), trong đó TP (True Positive) biểu thị cho số lượng trường hợp bạo lực, phát hiện đúng bởi mô hình, FP (False Positive) biểu thị cho số lượng trường hợp bình thường (không phải bạo lực) bị mô hình phân loại nhầm thành trường hợp bạo lực, TN (True Negative) biểu thị cho số lượng trường hợp bình thường, phân loại đúng bởi mô hình, FN (False Negative) biểu thị cho số lượng trường hợp bạo lực mô hình không phát hiện.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{F1} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

3. Kết quả và thảo luận

Mô hình C-ViDNet được huấn luyện trên các bộ dữ liệu Hockey Fight (HF), Violence in Movies (M), Real Life Violence Situations (RLS), Violence Detection (AVD). Kết quả thực nghiệm được trình bày như trong Bảng 3 và trực quan trong Hình 6. Để cung cấp góc nhìn chi tiết hơn về hiệu suất của các mô hình, chúng tôi tiến hành so sánh hiệu suất của mô hình được đề xuất với các nghiên cứu khác trên cùng bộ dữ liệu. Các so sánh về độ chính xác (Accuracy) được trình bày trong các Bảng 4, 5, 6, 7 và so sánh về số lượng tham số của các mô hình được trực quan trong Hình 7.

Bảng 3. *Kết quả kiểm thử của C-ViDNet trên các bộ dữ liệu*

Bộ dữ liệu	Accuracy	F1-score	Tham số
HF	95.50%	90.78%	853.889
VM	97.50%	86.67%	
RLVS	95.25%	81.43%	
AVD	92.85%	94.81	

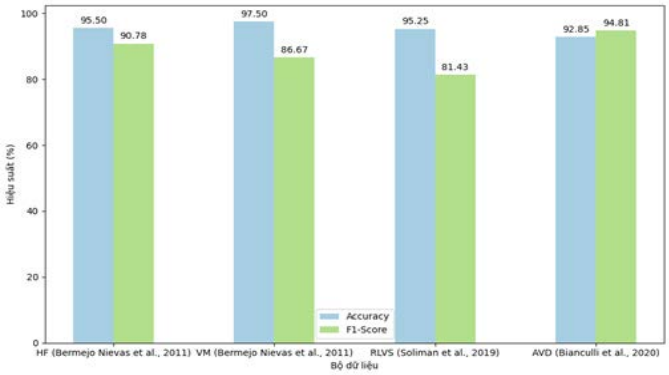
Bảng 4. *So sánh kết quả kiểm thử của C-ViDNet với các mô hình trên tập test của bộ HF*

Nghiên cứu	Accuracy	Tham số
(Halder & Chatterjee, 2020)	99.27%	1.8M
(Huszár et al., 2023)	97.50%	2.98M
(Li et al., 2019)	98.30%	7.4M
C-ViDNet	95.50%	853.889

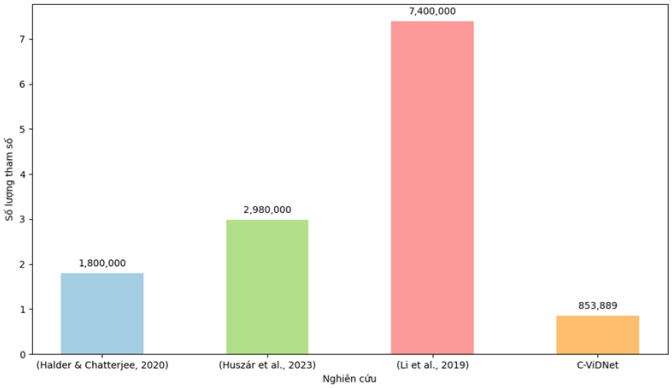
Mặc dù C-ViDNet không đạt được độ chính xác cao nhất so với các nghiên cứu trước đó, mô hình lại nổi bật nhờ số lượng tham số ít, mang đến một sự đánh đổi hợp lí giữa hiệu suất và hiệu quả tính toán. Với chỉ 853.889 tham số, mô hình nhỏ gọn hơn rất nhiều so với (Li et al., 2019) (7.4 triệu tham số), (Huszár et al., 2023) (2.98 triệu tham số) và (Halder & Chatterjee, 2020) (1.8 triệu tham số). Điều này đặc biệt quan trọng trong các hệ thống thực

tế, nơi tài nguyên phần cứng thường bị giới hạn, như các thiết bị nhúng hoặc các ứng dụng thời gian thực. Việc giảm số lượng tham số không chỉ giúp tiết kiệm năng lượng mà còn giảm thời gian xử lí, từ đó nâng cao khả năng triển khai trên diện rộng.

Dù có độ chính xác thấp hơn, như trên tập dữ liệu HF (95.50%) so với (Halder & Chatterjee, 2020) (99.27%), C-ViDNet vẫn duy trì được hiệu suất phân loại ở mức cao. Điều này cho thấy rằng, mặc dù có sự đánh đổi về độ chính xác, mô hình vẫn đảm bảo khả năng hoạt động hiệu quả trong các ứng dụng không yêu cầu độ chính xác tuyệt đối. Đặc biệt, trên các tập dữ liệu như VM và AVD, C-ViDNet đạt độ chính xác lần lượt là 97.50% và 92.85%, cho thấy mô hình vẫn có tính ứng dụng cao trong nhiều tình huống thực tế. C-ViDNet đã đạt được F1-score ấn tượng trên các bộ dữ liệu, với 90.78% trên HF, 86.67% trên MV, và 81.43% trên RLVS. Kết quả này làm nổi bật hiệu quả phân loại chính xác của mô hình mặc dù không sử dụng một kiến trúc phức tạp.



Hình 6. *Kết quả kiểm thử của mô hình C-ViDNet trên các bộ dữ liệu*



Hình 7. *Số lượng tham số của các mô hình*

Bảng 5. *So sánh kết quả kiểm thử của C-ViDNet với các mô hình trên bộ VM*

Nghiên cứu	Accuracy	Tham số
(Halder & Chatterjee, 2020)	100%	1.8M
(Huszár et al., 2023)	97.50%	2.98M
(Li et al., 2019)	100%	7.4M
C-ViDNet	97.50%	853.889

Sự đánh đổi giữa độ chính xác và số lượng tham số này là một hướng đi hợp lí trong thiết kế mô hình, đặc biệt khi xét đến các bài toán thực tế. Với số lượng tham số thấp, C-ViDNet phù hợp với các ứng dụng cần tính toán nhanh, hiệu quả và tiết kiệm tài nguyên, chẳng hạn như các hệ thống giám sát an ninh thời gian thực.

Bảng 6. *So sánh kết quả kiểm thử của C-ViDNet với các mô hình trên bộ RLVS*

Nghiên cứu	Accuracy	Tham số
(Huszár et al., 2023)	96.70%	2.98M
(Abdali, 2021)	96.25%	-
C-ViDNet	95.25%	853.889

Bảng 7. *So sánh kết quả kiểm thử của C-ViDNet với các mô hình trên bộ AVD*

Nghiên cứu	**Accuracy**	**Tham số**
(Siddique et al., 2022)	88.89%	-
(Haque et al., 2022)	90.00%	371.009
C-ViDNet	92.85%	853.889

Hình 8. *Kết quả của C-ViDNet với các tình huống bạo lực (trái), không bạo lực (phải)*

Hình 9. *Một số tình huống mô hình C-ViDNet bị nhầm lẫn*

Để có góc nhìn thực tế hơn, chúng tôi cũng tiến hành kiểm tra hiệu quả dự đoán của mô hình trên một số mẫu dữ liệu (Hình 8, 9) được chúng tôi tự thu thập từ thực tế và từ bộ dữ liệu AIRTLab (Sernani et al., 2021). Dù mô hình đã có thể phát hiện được hầu hết các

tình huống bạo lực, tuy nhiên vẫn còn tồn tại một số hạn chế. Chẳng hạn như không phát hiện được bạo lực trong các tình huống bị che khuất, nhầm lẫn giữa hành vi bạo lực với các hành động gần gũi thân mật, hay dự đoán là không bạo lực trong các tình huống sử dụng vũ khí (gậy) khi mà khoảng cách giữa những người trong cảnh cách xa nhau.

Ngoài ra, dù C-ViDNet có số lượng tham số huấn luyện ít hơn đáng kể so với các nghiên cứu trước đây, nhưng do quá trình xử lí và trích xuất đặc trưng phức tạp khiến mô hình gặp hạn chế lớn ở khả năng phát hiện bạo lực học đường theo thời gian thực (real-time). Hơn nữa, mô hình hiện tại chỉ tập trung khai thác các đặc trưng hình ảnh mà bỏ qua các phương thức khác, do đó có thể bỏ qua một số trường hợp bạo lực tiềm ẩn. Việc mở rộng mô hình theo hướng đa phương thức, tích hợp thêm tín hiệu âm thanh, là một hướng đi tiềm năng mà chúng tôi dự kiến triển khai trong các nghiên cứu tiếp theo.

4. Kết luận và kiến nghị

Trong nghiên cứu này, một hướng tiếp cận hiệu quả sử dụng biểu diễn khung xương dưới dạng heatmap đã được đề xuất. Mô hình C-ViDNet của chúng tôi đã giảm thiểu một số các đặc trưng không cần thiết như nền, trang phục và kích thước khung hình. Cách tiếp cận này đã giúp giảm độ phức tạp của mô hình so với các phương pháp khác dựa trên chuỗi khung hình video. Mặc dù mô hình đã đạt được khả năng nhận diện phần lớn các tình huống bạo lực, nhưng vẫn tồn tại một số hạn chế đáng chú ý. Cụ thể, C-ViDNet gặp khó khăn trong việc phát hiện bạo lực khi các hành động bị che khuất, dễ nhầm lẫn giữa hành vi bạo lực và các cử chỉ thân mật, hoặc dự đoán sai là không có bạo lực trong các tình huống sử dụng vũ khí (như gậy) khi khoảng cách giữa các đối tượng trong cảnh quá xa nhau. Trong tương lai, chúng tôi sẽ tiếp tục thử nghiệm các phương pháp cải tiến như nâng cao chất lượng khung xương, mở rộng huấn luyện trên nhiều góc nhìn khác nhau và tăng cường đặc trưng của bối cảnh xung quanh.

❖ ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

❖ ***Lời cảm ơn:*** *Nghiên cứu này được tài trợ bởi Nguồn ngân sách khoa học và công nghệ Trường Đại học Sư phạm Thành phố Hồ Chí Minh trong đề tài mã số CS.2024.19.32.*

TÀI LIỆU THAM KHẢO

Abdali, A. R. (2021). Data efficient video transformer for violence detection. *2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)* (pp. 195-199). IEEE. https://doi.org/10.1109/COMNETSAT53002.2021.9530829

Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., & Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II* (pp. 332-339). Springer. https://doi.org/10.1007/978-3-642-22993-9_42

Bianculli, M., Falcionelli, N., Sernani, P., Tomassini, S., Contardo, P., Lombardi, M., & Dragoni, A. F. (2020). A dataset for automatic violence detection in videos. *Data in Brief, 33*, Article 106587. https://doi.org/10.1016/j.dib.2020.106587

Divya, A., Lakshmi, D. S., Niveditha, P. L. N., Sri, P. S. N. S., Rohith, V., & Tati, V. B. (2024). Dual-stage deep learning framework for effective public physical violence detection. In *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 637-642). IEEE. https://doi.org/10.1109/CSNT60213.2024.10545798

Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2969-2978). https://doi.org/10.1109/CVPR52688.2022.00298

Ghalley, A., Abdelsalam, A., Dombola, W., & Choudhary, M. S. (2024). Violence detection in automated surveillance using CNN. In *2024 4th International Conference on Intelligent Technologies (CONIT)* (pp. 1-6). IEEE. https://doi.org/10.1109/CONIT61985.2024.10626390

Government. (2017). *Nghị định số 80/2017/NĐ-CP: Quy định về môi trường giáo dục an toàn, lành mạnh, thân thiện, phòng, chống bạo lực học đường* [*Decree No. 80/2017/ND-CP: Regulations on a safe, healthy, and friendly educational environment and prevention of school violence*].

Halder, R., & Chatterjee, R. (2020). CNN-BiLSTM model for violence detection in smart surveillance. *SN Computer Science, 1*(4), Article 201. https://doi.org/10.1007/s42979-020-00324-9

Huszár, V. D., Adhikarla, V. K., Négyesi, I., & Krasznay, C. (2023). Toward fast and accurate violence detection for automated video surveillance applications. *IEEE Access, 11*, 18772-18793. https://doi.org/10.1109/ACCESS.2023.3245521

Islam, Z., Rukonuzzaman, M., Ahmed, R., Kabir, M. H., & Farazi, M. (2021). Efficient two-stream network for violence detection using separable convolutional LSTM. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9533586

Khan, M., et al. (2024). Action knowledge graph for violence detection using audiovisual features. In *2024 IEEE International Conference on Consumer Electronics (ICCE)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICCE59016.2024.10444158

Kumar, M., Patel, A. K., Biswas, M., & Shitharth, S. (2023). Attention-based bidirectional long short-term memory for abnormal human activity detection. *Scientific Reports, 13*(1), Article 14442. https://doi.org/10.1038/s41598-023-14442-1

Li, J., Jiang, X., Sun, T., & Xu, K. (2019). Efficient violence detection using 3D convolutional neural networks. In *2019 16th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)* (pp. 1-8). IEEE. https://doi.org/10.1109/AVSS.2019.8909883

Parui, S. K., Biswas, S. K., Das, S., Chakraborty, M., & Purkayastha, B. (2023). An efficient violence detection system from video clips using ConvLSTM and keyframe extraction. In *2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)* (pp. 1-5). IEEE. https://doi.org/10.1109/IEMECON123456

Rutherford, A., Zwi, A. B., Grove, N. J., & Butchart, A. (2007). Violence: A glossary. *Journal of Epidemiology & Community Health, 61*(8), 676-680. https://doi.org/10.1136/jech.2005.043711

Santos, F., Durães, D., Marcondes, F. S., Hammerschmidt, N., Lange, S., Machado, J., & Novais, P. (2021). In-car violence detection based on the audio signal. In *Intelligent Data Engineering and Automated Learning – IDEAL 2021* (Vol. 13113, pp. 525-535). Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-030-91608-4_43

Sernani, P., Falcionelli, N., Tomassini, S., Contardo, P., & Dragoni, A. F. (2021). Deep learning for automatic violence detection: Tests on the AIRTLab dataset. *IEEE Access, 9*, 160580-160595. https://doi.org/10.1109/ACCESS.2021.3051347

Siddique, L. A., Junhai, R., Reza, T., Khan, S. S., & Rahman, T. (2022). Analysis of real-time hostile activity detection from spatiotemporal features using time distributed deep CNNs, RNNs, and attention-based mechanisms. In *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)* (pp. 1-6). IEEE. https://doi.org/10.1109/IPAS56160.2022.00016

Soliman, M. M., Kamal, M. H., El-Massih Nashed, M. A., Mostafa, Y. M., Chawky, B. S., & Khattab, D. (2019). Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)* (pp. 80-85). IEEE. https://doi.org/10.1109/ICICIS46948.2019.9014714

UNICEF. (2021). Protecting children from violence in school. *Retrieved September 6, 2024, from* https://www.unicef.org/protection/violence-against-children-in-school

World Health Organization, Regional Office for the Eastern Mediterranean. (2024). *Violence*. Retrieved September 30, 2024, from https://www.emro.who.int/health-topics/violence/index.html

Wu, P., Liu, X., & Liu, J. (2023). Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia, 25*, 1674-1685. https://doi.org/10.1109/TMM.2022.3147369

Yildiz, A. M., Barua, P. D., Dogan, S., Baygin, M., Tuncer, T., Ooi, C. P., Fujita, H., & Acharya, U. R. (2023). A novel tree pattern-based violence detection model using audio signals. *Expert Systems with Applications, 224*, Article 120031. https://doi.org/10.1016/j.eswa.2023.120031

C-VIDNET: AMODEL FOR SUPPORTING VIOLENCE DETECTION IN SCHOOLS

Nguyen Viet Hung, Ta Cong Phi*, Le Tan Loc, Ngo Quang Khanh, Tran Thanh Nha

Ho Chi Minh City University of Education, Vietnam

**Corresponding Author: Ta Cong Phi – Email: tacongphi1@gmail.com*

Received: January 20, 2025; Revised: April 23, 2025; Accepted: April 24, 2025

ABSTRACT

School violence is a complex and concerning issue within the education systems of many countries worldwide, including Vietnam. Although various automatic violence detection models based on artificial intelligence have been developed, practical implementation remains challenging due to high complexity and computational costs. To address these limitations, our study proposes the development of a C-ViDNet (Campus Violence Detection Network) model for automatic school violence detection with a small number of parameters, aimed at enhancing detection capabilities and rapid response to violent incidents in educational environments. First, YOLOX is used to identify individuals appearing in the frame. Next, the poses of these individuals are extracted using HRNet and converted into 3D Heatmap Volumes, helping to reduce noise and eliminate unnecessary background elements. Then, a dual-stream architecture is implemented to learn features from the 3D Heatmap Volumes. One stream focuses on the spatial features of the poses, while the other monitors changes in human poses across frames. The results from C-ViDNet demonstrate the potential for developing automatic school violence detection models. This approach minimizes dependence on manual monitoring and enables timely responses to violent incidents, contributing to safer educational environments.

Keywords: behavior recognition; computer vision; image processing; school violence; violence; YOLO