ĐẠI HỌC HUẾ
TRƯỜNG ĐẠI HỌC KHOA HỌC

NGUYỄN DŨNG

NÂNG CAO HIỆU QUẢ NHẬN DẠNG ĐỐI TƯỢNG VÀ HÀNH ĐỘNG TỪ CAMERA GIÁM SÁT

TÓM TẮT LUẬN ÁN TIẾN SĨ
KHOA HỌC MÁY TÍNH

ĐẠI HỌC HUẾ
TRƯỜNG ĐẠI HỌC KHOA HỌC

NGUYỄN DŨNG

NÂNG CAO HIỆU QUẢ NHẬN DẠNG ĐỐI TƯỢNG VÀ HÀNH ĐỘNG TỪ CAMERA GIÁM SÁT

Chuyên ngành: **KHOA HỌC MÁY TÍNH**
Mã số: **9480101**

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học

Hướng dẫn 1: **PGS. TS. Hoàng Văn Dũng**
Hướng dẫn 2: **TS. Lê Văn Tường Lân**

MỞ ĐẦU

Luận án tập trung vào bài toán nâng cao hiệu quả nhận dạng đối tượng và hành động từ camera giám sát, một hướng nghiên cứu có ý nghĩa khoa học và thực tiễn cao trong bối cảnh các hệ thống camera ngày càng được triển khai rộng rãi tại đô thị, khu dân cư, cơ sở hạ tầng trọng yếu và môi trường công cộng. Khối lượng dữ liệu hình ảnh và video phát sinh rất lớn khiến việc quan sát và phân tích thủ công không còn khả thi, trong khi các tình huống nguy hiểm như bạo lực, sự xuất hiện của vũ khí hoặc các đối tượng bất thường thường diễn ra đột ngột, đòi hỏi hệ thống phải có khả năng hỗ trợ phát hiện chính xác và kịp thời.

Từ nhu cầu đó, luận án xác định các nhóm thách thức chính cần giải quyết. Thứ nhất là dữ liệu mất cân bằng, do các trường hợp đối tượng hiếm hoặc bất thường thường xuất hiện ít so với dữ liệu của các lớp bình thường. Thứ hai là phát hiện đối tượng nhỏ trong ảnh và video giám sát, nhất là trong bối cảnh góc nhìn rộng, độ phân giải hạn chế hoặc đối tượng bị che khuất. Thứ ba là phát hiện hành động bạo lực trong video, nơi hệ thống cần cô lập vùng chuyển động có ý nghĩa và mô hình hóa sự biến đổi của hành vi theo thời gian.

Mục tiêu tổng quát của luận án là nghiên cứu và đề xuất các phương pháp nhằm nâng cao hiệu quả nhận dạng đối tượng và hành động từ camera giám sát. Để hiện thực hóa mục tiêu này, luận án được triển khai theo các hướng nội

dung bổ trợ cho nhau: xây dựng cơ chế học hiệu quả trên dữ liệu mất cân bằng; cải thiện khả năng phát hiện đối tượng nhỏ; phát hiện hành động bạo lực trong video từ camera giám sát theo đặc tính không gian–thời gian; và đánh giá thực nghiệm trên các tập dữ liệu chuẩn.

TỔNG QUAN VẤN ĐỀ NGHIÊN CỨU

Luận án tiếp cận bài toán nhận dạng đối tượng và hành động từ camera giám sát theo hai hướng chính: nhận dạng theo đối tượng và nhận dạng theo hành động. Hướng thứ nhất tập trung phát hiện sự xuất hiện của các đối tượng quan tâm trong khung hình và đặc biệt là các đối tượng có kích thước nhỏ. Hướng thứ hai hướng đến nhận diện các chuỗi chuyển động, tương tác và diễn biến theo thời gian của các đối tượng, chẳng hạn như hành động bạo lực, xâm nhập trái phép hoặc các tình huống bất thường trong đám đông. Hai hướng này có quan hệ bổ trợ, bởi bước phát hiện và theo dõi đối tượng có thể cung cấp các vùng quan tâm để biểu diễn chuyển động tập trung hơn vào con người. Trong phạm vi bài toán phát hiện bạo lực, luận án sử dụng phát hiện và theo dõi người để xác định vùng không gian của từng cá thể, sau đó mô hình hóa chuỗi đặc trưng thống kê của dòng quang học theo thời gian.

Trên phương diện phương pháp, luận án kế thừa các nền tảng học máy, học sâu và thị giác máy tính hiện đại, bao gồm các mô hình học có giám sát, học trên dữ liệu mất cân bằng, các kiến trúc CNN, RNN, Transformer và cơ chế chú ý, cùng các mô hình phát hiện đối tượng một giai đoạn, hai giai đoạn và dựa trên Transformer. Đối với dữ liệu video, luận án khảo sát các kỹ thuật mô hình hóa chuyển động và lựa chọn quy trình kết hợp phát hiện người, theo dõi đa đối tượng và dòng quang học. Từ quá trình tổng quan, luận án

xác định các thách thức nổi bật gồm mất cân bằng dữ liệu, phát hiện đối tượng nhỏ, tính bền vững của mô hình trong môi trường công cộng và khả năng nhận dạng hành động bất thường theo đặc tính không gian–thời gian. Đây là cơ sở để xây dựng các chương nghiên cứu tiếp theo theo hướng kết hợp giữa đóng góp phương pháp luận nền tảng và các bài toán ứng dụng chuyên biệt.

Trên cơ sở đó, luận án đã đạt được một số kết quả chính có ý nghĩa cả về phương pháp luận lẫn thực nghiệm. Thứ nhất, luận án đề xuất cơ chế gán trọng số thích nghi động trong hàm mất mát nhằm nâng cao hiệu quả học trên dữ liệu mất cân bằng, qua đó cải thiện chất lượng nhận dạng trên các lớp hiếm và tạo một công cụ nền tảng có khả năng tích hợp vào nhiều bài toán thị giác máy tính khác nhau. Thứ hai, luận án phát triển mô hình phát hiện đối tượng nhẹ có tăng cường chú ý đa tỉ lệ nhằm nâng cao khả năng phát hiện đối tượng nhỏ trong ảnh và video giám sát, đồng thời bảo đảm hiệu quả tính toán cho các nền tảng tài nguyên hạn chế. Thứ ba, đối với bài toán nhận dạng theo hành động, luận án đề xuất quy trình phát hiện bạo lực lấy con người làm trung tâm, trong đó dòng quang học được tính trong vùng theo dõi của từng cá thể và được tóm tắt bằng các đặc trưng thống kê trên chuỗi thời gian ngắn.

Phần tổng quan đã hệ thống hóa các hướng nghiên cứu, các nhóm phương pháp chủ yếu và những thách thức đặc trưng của bài toán, đồng thời làm rõ khoảng trống nghiên cứu và định vị các đóng góp chính của luận án. Trên cơ sở đó, các chương tiếp theo được triển khai theo một cấu trúc

logic, trong đó mỗi chương tập trung giải quyết một khía cạnh trọng tâm của bài toán tổng thể nhưng vẫn bảo đảm tính liên kết thống nhất với mục tiêu nâng cao hiệu quả nhận dạng đối tượng bất thường trong điều kiện giám sát thực tế.

NỘI DUNG, KẾT QUẢ NGHIÊN CỨU

Chương 1. Tổng quan về nhận dạng đối tượng và hành động từ camera giám sát

Chương 1 trình bày cơ sở lý thuyết và bối cảnh nghiên cứu cho toàn bộ luận án. Nội dung chương tập trung làm rõ đặc trưng của dữ liệu ảnh và video giám sát, sự khác biệt giữa thông tin không gian và thông tin thời gian, cùng các nền tảng phương pháp chủ yếu của học máy, học sâu và thị giác máy tính hiện đại. Bên cạnh đó, chương tổng hợp các hướng tiếp cận tiêu biểu trong phát hiện đối tượng, nhận dạng hành động, các tiêu chí đánh giá thường dùng và những bộ dữ liệu chuẩn có liên quan đến phát hiện đối tượng và phát hiện hành động bạo lực từ camera giám sát.

Một nội dung trọng tâm của chương là phân tách bài toán thành hai hướng tiếp cận chính: nhận dạng theo đối tượng và nhận dạng theo hành động. Hướng thứ nhất tập trung vào việc phát hiện các đối tượng hiếm gặp, nguy hiểm hoặc không mong muốn trong khung hình, như vũ khí, vật thể lạ hoặc các mục tiêu có kích thước nhỏ. Hướng thứ hai nhấn mạnh việc phát hiện các chuỗi chuyển động, tương tác và diễn biến hành động theo thời gian, chẳng hạn như hành động bạo lực, xâm nhập trái phép hoặc các tình huống bất thường trong đám đông. Cách phân tách này giúp xác định rõ phạm vi nghiên cứu của từng chương, đồng thời làm sáng tỏ mối liên hệ giữa các hướng nghiên cứu thành phần trong một mục tiêu chung của luận án.

Về phương diện học thuật, Chương 1 có vai trò xác lập khung vấn đề nghiên cứu, chỉ ra các thách thức cốt lõi và định vị các hướng tiếp cận mà luận án lựa chọn. Trên cơ sở đó, chương làm rõ mối liên hệ giữa các bài toán dữ liệu mất cân bằng, phát hiện đối tượng nhỏ, phát hiện đối tượng trong môi trường công cộng và phát hiện hành động bạo lực trong video. Đây là cơ sở để luận án triển khai các đóng góp theo hướng kết hợp giữa phương pháp luận nền tảng và các bài toán ứng dụng chuyên biệt, bảo đảm tính thống nhất về mục tiêu nghiên cứu và tính liên kết giữa các chương.

Chương 2. Phương pháp học trên dữ liệu mất cân bằng

Chương 2 tập trung giải quyết bài toán dữ liệu mất cân bằng thông qua việc đề xuất cơ chế gán trọng số thích nghi động trong hàm mất mát, ký hiệu là DFL. Ý tưởng cốt lõi của phương pháp là điều chỉnh đóng góp của từng mẫu trong quá trình huấn luyện dựa trên phân bố lớp và độ khó của mẫu, nhằm giảm sự chi phối của lớp đa số và duy trì ảnh hưởng của các mẫu dương hiếm. Phương pháp được khảo sát không chỉ trên phân loại ảnh mà còn trên phát hiện đối tượng và phân đoạn ảnh.

Về mặt phương pháp, DFL được xây dựng như một biến thể thích nghi của Focal Loss, trong đó tham số tập trung được cập nhật động theo từng epoch. Cụ thể, hàm mất mát được viết dưới dạng:

$$DFL(p_t) = -\alpha(1 - p_t)^{\gamma_t} \log(p_t), \tag{1}$$

trong đó tham số γ_t được xác định theo công thức:

$$\gamma_t = \gamma_0 + (\gamma_{\text{final}} - \gamma_0) \left(\frac{t-1}{T-1} \right)^p . \tag{2}$$

Trong đó, p_t là xác suất dự đoán đúng lớp, α là hệ số cân bằng lớp, γ_0 là giá trị khởi tạo, γ_{final} là giá trị ở cuối quá trình huấn luyện, t là epoch hiện tại, T là tổng số epoch và p là tham số điều khiển tốc độ thay đổi. Nhờ cơ chế này, mô hình có thể điều chỉnh mức độ ưu tiên dành cho các mẫu khó theo từng giai đoạn học, thay vì duy trì một mức tập trung cố định trong toàn bộ quá trình huấn luyện.

Trong toàn bộ các thí nghiệm, DFL sử dụng chung cấu hình $\gamma_0 = 0$, $\gamma_{\text{final}} = 5$, $\alpha = 0.25$ và $p = 1$, không tinh chỉnh riêng theo từng tác vụ. Kết quả trên ISIC2019 cho thấy DFL cải thiện các chỉ số phân loại qua nhiều kiến trúc. Đối với phát hiện đối tượng, tác động tích cực thể hiện rõ hơn ở mAP@50 và mAP_s, nhưng không tăng đồng thời mọi chỉ số; một số cấu hình giảm nhẹ ở mAP@75 hoặc mAP_m. Trong bài toán phân đoạn ảnh, việc thay Cross-Entropy bằng DFL trong nhánh phân loại của RPN cải thiện kết quả trên cả hai cấu hình Mask R-CNN được khảo sát. Các kết quả cho thấy DFL là một cơ chế gán trọng số động có khả năng tích hợp vào nhiều tác vụ, đồng thời cần được diễn giải theo từng chỉ số và điều kiện thực nghiệm cụ thể.

Ngoài giá trị thực nghiệm, Chương 2 còn cho thấy tầm quan trọng của thiết kế hàm mất mát trong việc chi phối hành vi học của mô hình. Thay vì chỉ tập trung vào

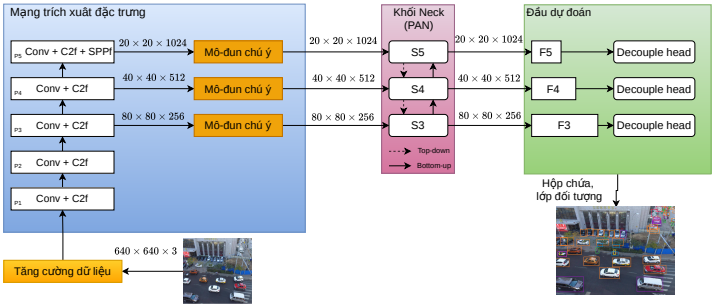
thay đổi kiến trúc, nghiên cứu nhấn mạnh rằng việc điều chỉnh cơ chế tối ưu ở mức hàm mất mát có thể tạo ra cải thiện đáng kể mà vẫn giữ được tính linh hoạt khi tích hợp với nhiều họ mô hình khác nhau. Đây là một đóng góp có ý nghĩa vì nó cung cấp một công cụ nền tảng để hỗ trợ các chương ứng dụng sau của luận án.

Chương 3. Phát hiện đối tượng có kích thước nhỏ

Chương 3 đề xuất LWMA-Net, một kiến trúc phát hiện đối tượng nhẹ có tăng cường cơ chế chú ý đa tỉ lệ, được thiết kế cho bài toán phát hiện đối tượng nhỏ trong ảnh UAV. Mô hình hướng tới giải quyết các khó khăn điển hình của dữ liệu chụp từ trên cao như kích thước mục tiêu nhỏ, nền cảnh phức tạp, mật độ đối tượng cao và yêu cầu triển khai trên phần cứng hạn chế. Điểm nổi bật của mô hình là cơ chế tinh chỉnh và hợp nhất đặc trưng dựa trên chú ý đa tỉ lệ, cho phép nhấn mạnh các vùng đặc trưng quan trọng đối với mục tiêu nhỏ hoặc bị che khuất.

Bên cạnh thiết kế kiến trúc, chương còn phân tích vai trò của từng thành phần trong LWMA-Net thông qua các thực nghiệm cắt bỏ và trực quan hóa. Các kết quả cho thấy cơ chế chú ý không chỉ làm tăng độ chính xác phát hiện mà còn giúp mô hình tập trung tốt hơn vào vùng đối tượng thực, đặc biệt trong các trường hợp mục tiêu có kích thước rất nhỏ, độ tương phản thấp hoặc bị nhiễu nền mạnh. Điều này làm rõ tính hợp lý của lựa chọn kiến trúc theo hướng nhẹ nhưng vẫn có khả năng mô hình hóa đặc trưng đa tỉ lệ hiệu quả.

Chiến lược huấn luyện của LWMA-Net sử dụng tăng cường dữ liệu, bộ tối ưu AdamW và lịch giảm tốc độ học dạng cosine. Trên bộ dữ liệu VisDrone2019, mô hình đạt 29.0% mAP và 47.4% mAP@50, đồng thời cải thiện rõ rệt hiệu năng trên đối tượng nhỏ so với mô hình cơ sở. Thí nghiệm tích hợp DFL của Chương 2 vào LWMA-Net tiếp tục nâng kết quả lên 29.5% mAP, 48.3% mAP@50, 30.1% mAP@75, 21.3% mAP_s, 40.0% mAP_m và 41.4% mAP_l. Do DFL chỉ tác động trong giai đoạn huấn luyện, cấu hình tích hợp không làm thay đổi số tham số, FLOPs và tốc độ suy luận của LWMA-Net. Kết quả này cung cấp bằng chứng thực nghiệm về tính bổ trợ giữa cơ chế gán trọng số động và cơ chế chú ý đa tỉ lệ.



Hình 3.1. Kiến trúc mô hình đề xuất LWMA-Net

Hình 3.1 minh họa cấu trúc tổng thể của LWMA-Net, trong đó các thành phần trích xuất đặc trưng, hợp nhất đa tỉ lệ và mô-đun chú ý được tổ chức theo hướng ưu tiên phát hiện mục tiêu nhỏ nhưng vẫn giữ số lượng tham số ở mức thấp. Việc đưa hình kiến trúc vào bản tóm tắt giúp làm rõ

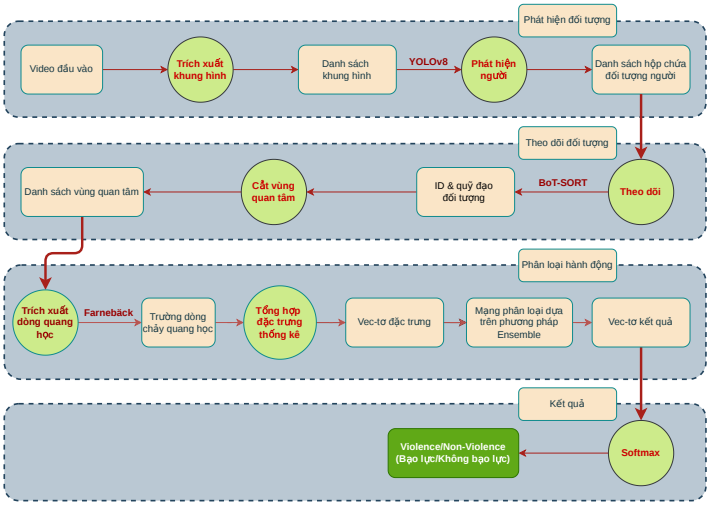
trực quan ý tưởng thiết kế trọng tâm của chương, đồng thời phản ánh đúng cấu trúc phương pháp đã trình bày trong luận án.

Chương 4. Phát hiện hành động bạo lực trong video theo đặc tính không gian–thời gian

Chương 4 trình bày cách tiếp cận phát hiện hành động bạo lực lấy con người làm trung tâm. Phát hiện và theo dõi người được sử dụng để xác định vùng quan tâm của từng cá thể; dòng quang học sau đó được tính trong các vùng này và tóm tắt bằng 25 đặc trưng thống kê về độ lớn và hướng chuyển động trên cửa sổ 15 khung hình. Ảnh RGB và đặc trưng ngoại hình không được đưa trực tiếp vào bộ phân loại. Vì vậy, thuật ngữ không gian–thời gian trong chương phản ánh việc biểu diễn chuyển động trong vùng không gian của từng người và mô hình hóa sự biến đổi của biểu diễn đó theo thời gian.

Để tránh rò rỉ do các cửa sổ liền kề chồng lấn, dữ liệu được chia theo nhóm với tỉ lệ huấn luyện/xác thực/kiểm tra là 70/10/20. Đối với AIRTLab, hai video cùng cảnh được quay từ hai camera được xếp vào cùng một nhóm; đối với Movies Fight, mỗi video là một nhóm độc lập. Tất cả cửa sổ và chuỗi theo dõi của cùng nhóm chỉ thuộc một tập, và quá trình kiểm tra tự động xác nhận không có chồng lấn nguồn giữa các tập. Tám mô hình phân loại và một mô hình ensemble được đánh giá trên ba hạt giống 42, 1 và 2; kết quả được báo cáo dưới dạng trung bình $\pm$ độ lệch chuẩn ở cả mức cửa sổ và mức video.

Trên AIRTLab với bước trượt 1, ensemble gồm 3D-CNN, CNN–LSTM và Transformer đạt F1-macro $64.95 \pm 1.28\%$ và accuracy mức video $85.71 \pm 2.86\%$. Với bước trượt 5, các giá trị tương ứng là $65.13 \pm 1.70\%$ và $80.48 \pm 2.97\%$. Trên Movies Fight, 3D-CNN đạt F1-macro cao nhất $89.22 \pm 3.88\%$; XGBoost đạt accuracy mức video cao nhất $98.33 \pm 1.44\%$; ensemble đạt recall cao nhất $90.01 \pm 3.55\%$. Kết quả cho thấy hiệu năng phụ thuộc đáng kể vào đặc điểm miền dữ liệu và không có một mô hình duy nhất vượt trội trên mọi chỉ số.



Hình 4.1. Kiến trúc tổng thể của hệ thống

Hình 4.1 thể hiện chuỗi xử lý chính của hệ thống, từ phát hiện người, theo dõi đối tượng, trích xuất dòng quang học, tổng hợp đặc trưng thống kê đến phân loại hành động. Phân tích hiệu quả tính toán trên GPU NVIDIA RTX 3080 10 GB cho thấy độ trễ của các bộ phân loại đều dưới 11 ms

cho một cửa sổ; tuy nhiên, chi phí chính của toàn hệ thống nằm ở giai đoạn phát hiện–theo dõi và trích xuất dòng quang học. Đây là điểm cần tiếp tục tối ưu khi hướng đến xử lý thời gian thực.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận

Luận án đã nghiên cứu và đề xuất các phương pháp nhằm nâng cao hiệu quả nhận dạng đối tượng và hành động từ camera giám sát, tập trung vào việc giải quyết ba thách thức cốt lõi: dữ liệu mất cân bằng, phát hiện đối tượng nhỏ, và nhận dạng hành động bạo lực trong ảnh/video từ camera giám sát. Các kết quả chính đạt được bao gồm:

Thứ nhất, luận án đề xuất DFL, một chiến lược gán trọng số động thông qua việc tăng tham số tập trung γ theo tiến trình huấn luyện. Phương pháp sử dụng một cấu hình chung cho phân loại, phát hiện và phân đoạn, qua đó cho thấy khả năng tích hợp vào nhiều tác vụ thị giác máy tính. Kết quả thực nghiệm xác nhận xu hướng cải thiện, đặc biệt đối với các mẫu khó và đối tượng nhỏ, đồng thời cũng chỉ ra tác động của DFL không đồng đều trên mọi chỉ số.

Thứ hai, luận án phát triển LWMA-Net, một mô hình chú ý đa tỉ lệ nhẹ cho bài toán phát hiện đối tượng nhỏ trong ảnh UAV. Mô hình đạt 29.0% mAP và 47.4% mAP@50 trên VisDrone2019. Khi tích hợp DFL, kết quả tăng lên 29.5% mAP và 48.3% mAP@50 mà không thay đổi độ phức tạp suy luận. Thí nghiệm này cung cấp bằng chứng trực tiếp về tính liên kết và bổ trợ giữa đóng góp của Chương 2 và Chương 3.

Thứ ba, luận án đề xuất khung phát hiện hành động bạo lực dựa trên biểu diễn chuyển động theo vùng người. Đầu

vào của bộ phân loại là chuỗi đặc trưng thống kê của dòng quang học, không bao gồm nhánh đặc trưng RGB độc lập. Quy trình chia dữ liệu theo nhóm loại bỏ chồng lấn nguồn giữa các tập; thực nghiệm trên AIRTLab và Movies Fight với ba hạt giống cung cấp đánh giá thận trọng hơn về khả năng tổng quát hóa. Kết quả đồng thời cho thấy bước phân loại có độ trễ thấp, trong khi phát hiện–theo dõi và trích xuất dòng quang học là điểm nghẽn tính toán chính.

Tóm lại, luận án đã hoàn thành các mục tiêu nghiên cứu đề ra và đóng góp các giải pháp cho ba khía cạnh bổ trợ của bài toán nhận dạng đối tượng và hành động từ camera giám sát. Các kết quả cho thấy tiềm năng ứng dụng của các phương pháp đề xuất, đồng thời việc triển khai trong hệ thống thực tế vẫn cần được kiểm chứng thêm trên dữ liệu giám sát đa dạng và trong các điều kiện vận hành khác nhau. Bên cạnh đó, có thể thấy các chương của luận án không được tổ chức theo dạng phụ thuộc tuần tự, trong đó kết quả của chương trước phải được tích hợp thực nghiệm trực tiếp vào chương sau, mà theo hướng giải quyết các khía cạnh bổ sung của cùng một bài toán tổng thể. Cụ thể, Chương 2 đề xuất một cơ chế học có tính tổng quát cho dữ liệu mất cân bằng, trong khi Chương 3, Chương 4 phát triển các giải pháp chuyên biệt cho các bối cảnh ứng dụng khác nhau của nhận dạng đối tượng và hành động từ camera giám sát. Theo hướng này, đóng góp của Chương 2 có vai trò phương pháp luận nền tảng và đã được kiểm chứng bước đầu thông qua thí nghiệm tích hợp DFL vào LWMA-Net. Chương 4 được đánh giá theo một thiết lập riêng nhằm cô lập ảnh hưởng

của biểu diễn chuyển động và mô hình hóa thời gian.

2. Hướng phát triển

Dựa trên các kết quả đạt được và những hạn chế còn tồn tại, luận án đề xuất các hướng nghiên cứu và phát triển tiếp theo:

Mở rộng tích hợp DFL và LWMA-Net vào quy trình video: Luận án đã tích hợp DFL vào LWMA-Net và ghi nhận cải thiện trên VisDrone2019 mà không làm thay đổi kiến trúc suy luận. Hướng phát triển tiếp theo là đưa mô-đun phát hiện đối tượng và cơ chế học trên dữ liệu mất cân bằng này vào quy trình video, sau đó đánh giá ảnh hưởng đối với chất lượng chuỗi theo dõi và nhận dạng hành động.

Mở rộng đánh giá trên dữ liệu giám sát thực: AIRTLab và Movies Fight đều còn chịu ảnh hưởng của dàn dựng. Vì vậy, phương pháp phát hiện bạo lực cần được đánh giá thêm trên các bộ dữ liệu giám sát thực như RWF-2000 hoặc RLVS, đồng thời thực hiện thí nghiệm huấn luyện trên một miền và kiểm tra trên miền khác để khảo sát mức độ phụ thuộc miền.

Mở rộng sang các dạng đối tượng và hành động khác: Các hạn chế chính của các mô hình bao gồm sự phụ thuộc vào chất lượng dữ liệu huấn luyện, điều kiện ánh sáng, góc quay camera, mức độ che khuất, sự đa dạng của bối cảnh giám sát và chênh lệch phân bố giữa dữ liệu huấn luyện với dữ liệu thực tế. Do đó, khi triển khai trong môi trường mới, các mô hình cần được hiệu chỉnh, đánh giá lại và có thể cần bổ sung dữ liệu đặc thù của từng bối cảnh. Bên cạnh các bài

toán đã được nghiên cứu, một hướng phát triển tiếp theo là mở rộng phương pháp sang các dạng hành động khác như xâm nhập khu vực cấm, tụ tập bất thường, té ngã, cháy nổ, trộm cắp hoặc các hành vi nguy hiểm khác trong môi trường giám sát. Việc mở rộng này cần kết hợp phát hiện đối tượng, nhận dạng hành động và mô hình hóa ngữ cảnh để nâng cao khả năng ứng dụng của hệ thống trong các tình huống thực tế đa dạng.

Hoàn thiện kiểm định thống kê và phân tích độ tin cậy: Chương 4 đã được đánh giá trên ba hạt giống và báo cáo trung bình kèm độ lệch chuẩn. Đối với các thí nghiệm quy mô lớn ở Chương 2 và Chương 3, cần tiếp tục thực hiện nhiều lần chạy độc lập, bổ sung khoảng tin cậy và kiểm định ý nghĩa thống kê, đặc biệt khi chênh lệch giữa các cấu hình nhỏ hơn một điểm phần trăm.

Nghiên cứu học chuyển giao miền và học với ít dữ liệu: Phát triển các phương pháp thích ứng miền và học vài mẫu để mô hình có thể thích ứng nhanh chóng với môi trường mới và học được các loại bất thường mới từ số lượng mẫu hạn chế.

Tăng cường khả năng giải thích và hỗ trợ ra quyết định: Các kết quả trực quan hóa cần được phát triển theo hướng không chỉ minh họa vùng chú ý của mô hình mà còn hỗ trợ người vận hành hiểu được nguyên nhân tạo ra cảnh báo. Trong tương lai, có thể kết hợp bản đồ chú ý, quỹ đạo chuyển động, thông tin đối tượng và mô tả ngữ nghĩa ngắn gọn để cung cấp bằng chứng trực quan cho từng quyết định của hệ thống.

Khai thác dữ liệu không gán nhãn và thông tin bổ trợ: Một hướng nghiên cứu tiếp theo là tận dụng lượng lớn video giám sát chưa gán nhãn thông qua học tự giám sát, học tương phản hoặc tiền huấn luyện trên dữ liệu miền đích. Bên cạnh dữ liệu hình ảnh, các nguồn thông tin bổ trợ như thời gian, vị trí camera, vùng quan tâm hoặc thông tin ngữ cảnh của cảnh quan sát có thể được khai thác để tăng độ ổn định của hệ thống.

Tối ưu hóa quy trình triển khai và vận hành hệ thống: Đối với quy trình phát hiện bạo lực, cần ưu tiên tăng tốc giai đoạn phát hiện–theo dõi và trích xuất dòng quang học trên GPU, bởi đây là thành phần chiếm phần lớn thời gian xử lý. Bên cạnh đó, cần nghiên cứu cơ chế cảnh báo theo ngưỡng tin cậy, cập nhật mô hình sau triển khai và đánh giá tải khi xử lý đồng thời nhiều luồng camera.

Tích hợp với hạ tầng giám sát thông minh: Trong các ứng dụng thực tế, mô hình nhận dạng cần được kết nối với các thành phần quản lý camera, lưu trữ video, truy vấn sự kiện và cảnh báo thời gian thực. Do đó, một hướng phát triển quan trọng là xây dựng kiến trúc hệ thống theo mô-đun, cho phép mở rộng số lượng camera, phân quyền truy cập dữ liệu và tích hợp với các nền tảng quản lý an ninh hiện có.

DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ KẾT QUẢ NGHIÊN CỨU CỦA ĐỀ TÀI LUẬN ÁN

[CT1] **Nguyễn Dũng***, Hoàng Văn Dũng, Lê Văn Tường Lân (2024). *Khảo sát một số mô hình phát hiện đối tượng dựa vào học sâu*. Tạp chí Khoa học và Công nghệ Trường Đại học Khoa học, tập 23, số 1, trang 39–52 (Đã xuất bản).

[CT2] **Dung Nguyen**, Van-Dung Hoang*, Van-Tuong-Lan Le (2025). *Evaluation of the effectiveness of modern object detection models on spatial image data*. Hue University Journal of Science. DOI: 10.26459/hueunijtt.v134i2B.7862 (Đã xuất bản).

[CT3] **Nguyễn Dũng**, Hoàng Văn Dũng, Lê Văn Tường Lân* (2024). *Enhancing the DETR Object Detection Model by Integrating Linformer*. FAIR2024.
DOI: 10.15625/vap.2024.0210 (Đã xuất bản).

[CT4] **Dung Nguyen**, Van-Dung Hoang*, and Van-Tuong-Lan Le (2024). *V-DETR: Pure Transformer for End-to-End Object Detection*. ACIIDS 2024, LNCS 14796, Springer (Scopus Q2). DOI: 10.1007/978-981-97-4985-0_10 (Đã xuất bản).

[CT5] **Dung Nguyen**, Van-Dung Hoang*, Van-Tuong-Lan Le (2026). *Enhancing object detection efficiency with Transformers through multi-level feature integration*. Journal of Computer Science and Cybernetics. DOI: 10.15625/1813-9663/21114 (Đã xuất bản).

[CT6] **Dung Nguyen**, Van-Dung Hoang*, Van-Tuong-Lan Le

(2025). *Applying deep learning models for weapon detection in public environments through surveillance cameras*. Journal of Research and Development on ICT. DOI: 10.32913/mic-ict-research.v2024.n2.1318 (Đã xuất bản).

[CT7] **Dung Nguyen**, Van-Dung Hoang*, and Van-Tuong-Lan Le (2026). *Towards enhancing learning on imbalanced data: A novel adaptive weighting strategy*. Neurocomputing, vol 673 (SCIE, Q1, IF 6.7). DOI: 10.1016/j.neucom.2026.132886 (Đã xuất bản).

[CT8] **Dung Nguyen**, Van-Dung Hoang, and Van-Tuong-Lan Le* (2025). *A Lightweight Transformer Model For Real-Time Object Detection On Edge Devices*. ACIIDS 2025 (Scopus Q4). DOI: 10.1007/978-981-96-5884-8_16 (Đã xuất bản).

[CT9] **Dung Nguyen**, Van-Dung Hoang*, and Van-Tuong-Lan Le (2026). *A Lightweight Multi-Scale Attention Model for Small Object Detection in UAV Imagery*. IEEE Access (SCIE, Q1, IF 4.2). DOI: 10.1109/ACCESS.2026.3656179 (Đã xuất bản).

[CT10] **Dung Nguyen**, Van-Dung Hoang*, Van-Tuong-Lan Le (2026). *Tracklet-Driven Spatio-Temporal Representation Learning with Enhancing Optical Flow and Transformer feature for Violence Detection*. Engineering Applications of Artificial Intelligence (Đã gửi tạp chí).

HUE UNIVERSITY
UNIVERSITY OF SCIENCES

NGUYEN DUNG

Enhancing the Efficiency of Object and Action Recognition from Surveillance Cameras

SUMMARY OF DOCTORAL DISSERTATION
COMPUTER SCIENCE

HUE UNIVERSITY
UNIVERSITY OF SCIENCES

NGUYEN DUNG

Enhancing the Efficiency of Object and Action Recognition from Surveillance Cameras

Major: **COMPUTER SCIENCE**
Code: **9480101**

SUMMARY OF DOCTORAL DISSERTATION
COMPUTER SCIENCE

Scientific Supervisors

Supervisor 1: **Assoc. Prof. Dr. Hoang Van Dung**
Supervisor 2: **Dr. Le Van Tuong Lan**

INTRODUCTION

This dissertation focuses on enhancing the efficiency of object and action recognition from surveillance cameras, a research direction of substantial scientific and practical significance as camera systems are increasingly deployed in urban areas, residential zones, critical infrastructure, and public environments. The enormous volume of image and video data makes continuous manual monitoring and analysis impractical. Meanwhile, hazardous events such as violence, the presence of weapons, or other abnormal objects often occur suddenly and therefore require systems capable of supporting accurate and timely detection.

Based on these practical demands, the dissertation identifies three major groups of challenges. The first is data imbalance, because rare objects or abnormal cases generally appear much less frequently than normal classes. The second is small-object detection in surveillance images and videos, particularly under wide viewing angles, limited resolution, and occlusion. The third is violence detection in video, where the system must isolate meaningful motion regions and model behavioral changes over time.

The overall objective of the dissertation is to investigate and propose methods for enhancing the efficiency of object and action recognition from surveillance cameras. To achieve this objective, the research is organized into complementary directions: developing an effective learning mech-

anism for imbalanced data; improving small-object detection; detecting violent actions in surveillance video through spatio-temporal characteristics; and conducting experimental evaluations on benchmark datasets.

OVERVIEW OF THE RESEARCH PROBLEM

The dissertation approaches object and action recognition from surveillance cameras through two principal directions: object-oriented recognition and action-oriented recognition. The first direction focuses on detecting objects of interest in individual frames, especially small objects. The second aims to recognize motion sequences, interactions, and temporal developments, such as violent behavior, unauthorized intrusion, or abnormal crowd situations. These two directions are complementary because object detection and tracking can provide regions of interest that allow motion representations to focus more directly on people. For violence detection, the dissertation uses person detection and tracking to determine the spatial region of each individual and subsequently models temporal sequences of statistical optical-flow features.

Methodologically, the dissertation builds upon modern foundations in machine learning, deep learning, and computer vision, including supervised learning, learning from imbalanced data, CNN, RNN, Transformer and attention architectures, as well as one-stage, two-stage, and Transformer-based object detectors. For video data, the dissertation reviews motion-modeling techniques and adopts a pipeline that combines person detection, multi-object tracking, and optical flow. The literature review identifies several

prominent challenges, including data imbalance, small-object detection, model robustness in public environments, and abnormal-action recognition through spatio-temporal characteristics. These findings provide the basis for combining a foundational methodological contribution with specialized application-oriented studies in the subsequent chapters.

The dissertation achieves several major methodological and experimental results. First, it proposes a dynamic adaptive-weighting mechanism in the loss function to improve learning from imbalanced data, enhance recognition of rare classes, and provide a foundational tool that can be integrated into multiple computer-vision tasks. Second, it develops a lightweight object detector enhanced with multi-scale attention to improve small-object detection in surveillance imagery while maintaining computational efficiency for resource-constrained platforms. Third, for action-oriented recognition, it proposes a human-centric violence-detection pipeline in which optical flow is computed within the tracked region of each individual and summarized using statistical features over short temporal sequences.

The overview systematizes the main research directions, methodological groups, and task-specific challenges, while clarifying the research gaps and positioning the principal contributions of the dissertation. The subsequent chapters are therefore organized so that each addresses a central aspect of the overall problem while remaining aligned with the common objective of enhancing abnormal-object and ac-

tion recognition under practical surveillance conditions.

RESEARCH CONTENT AND RESULTS

Chapter 1. Overview of Object and Action Recognition from Surveillance Cameras

Chapter 1 presents the theoretical foundations and research context of the dissertation. It clarifies the characteristics of surveillance image and video data, the distinction between spatial and temporal information, and the principal methodological foundations of modern machine learning, deep learning, and computer vision. The chapter also reviews representative approaches to object detection and action recognition, commonly used evaluation criteria, and benchmark datasets related to object detection and violence detection from surveillance cameras.

A central aspect of the chapter is the division of the research problem into object-oriented and action-oriented recognition. Object-oriented recognition focuses on rare, dangerous, or unwanted objects in individual frames, such as weapons, unusual items, or small targets. Action-oriented recognition emphasizes motion sequences, interactions, and temporal developments, including violent behavior, unauthorized intrusion, and abnormal crowd situations. This division defines the scope of the subsequent chapters and clarifies the relationship among the component research directions within the common objective of the dissertation.

Academically, Chapter 1 establishes the research framework, identifies the core challenges, and positions the approaches selected in the dissertation. It clarifies the relationships among imbalanced learning, small-object detection, object detection in public environments, and violence detection in video. This foundation supports the combination of general methodological contributions and specialized application problems while maintaining coherence in the research objectives and interconnections among the chapters.

Chapter 2. Learning Method for Imbalanced Data

Chapter 2 addresses data imbalance by proposing a dynamic adaptive-weighting mechanism in the loss function, denoted DFL. The central idea is to adjust the contribution of each training sample according to the class distribution and sample difficulty, thereby reducing the dominance of majority classes while preserving the influence of rare positive samples. The method is evaluated not only for image classification but also for object detection and image segmentation.

DFL is formulated as an adaptive variant of Focal Loss in which the focusing parameter is dynamically updated at each epoch. The loss is defined as

$$DFL(p_t) = -\alpha(1 - p_t)^{\gamma_t} \log(p_t), \tag{1}$$

where γ_t is determined by

$$\gamma_t = \gamma_0 + (\gamma_{\text{final}} - \gamma_0) \left(\frac{t-1}{T-1} \right)^p . \tag{2}$$

Here, p_t is the predicted probability of the correct class, α is the class-balancing factor, γ_0 is the initial focusing value, γ_{final} is its final value, t is the current epoch, T is the total number of epochs, and p controls the rate of change. This mechanism allows the model to adjust its emphasis on difficult samples throughout training rather than maintaining a fixed focusing level.

All experiments use the common configuration $\gamma_0 = 0$, $\gamma_{\text{final}} = 5$, $\alpha = 0.25$, and $p = 1$, without task-specific tuning. On ISIC2019, DFL improves classification metrics across multiple architectures. In object detection, the positive effect is more apparent for mAP@50 and mAP_s, although not every metric improves simultaneously; several configurations exhibit slight decreases in mAP@75 or mAP_m. For image segmentation, replacing Cross-Entropy with DFL in the RPN classification branch improves performance for both evaluated Mask R-CNN configurations. These results demonstrate that DFL is a dynamic weighting mechanism applicable to multiple tasks, while its effects should be interpreted according to individual metrics and experimental conditions.

Beyond the empirical findings, Chapter 2 highlights the importance of loss-function design in governing model-learning behavior. Instead of relying exclusively on architectural changes, the study demonstrates that modifying the op-

timization mechanism at the loss-function level can improve performance while retaining flexibility across model families. DFL therefore provides a methodological foundation for the application-oriented chapters that follow.

Chapter 3. Small-Object Detection

Chapter 3 proposes LWMA-Net, a lightweight object-detection architecture enhanced with multi-scale attention for small-object detection in UAV imagery. The model addresses typical aerial-imagery challenges, including small target size, complex backgrounds, high object density, and deployment constraints on limited hardware. Its key design is a multi-scale attention-based feature-refinement and fusion mechanism that emphasizes informative features corresponding to small or occluded targets.

The chapter analyzes the contribution of each LWMA-Net component through ablation experiments and visualization. The results show that the attention mechanisms improve detection accuracy and enable the model to concentrate more effectively on actual object regions, particularly for extremely small targets, low-contrast objects, and scenes with strong background interference. These findings support the design choice of a lightweight architecture that still models multi-scale features effectively.

LWMA-Net is trained with data augmentation, the AdamW optimizer, and a cosine learning-rate schedule. On VisDrone2019, the model achieves 29.0% mAP and 47.4% mAP@50 while providing a marked improvement for small

objects over the baseline. Integrating DFL from Chapter 2 further increases performance to 29.5% mAP, 48.3% mAP@50, 30.1% mAP@75, 21.3% mAP_s, 40.0% mAP_m, and 41.4% mAP_l. Because DFL affects only the training stage, the integrated configuration preserves the parameter count, FLOPs, and inference speed of LWMA-Net. This experiment provides direct empirical evidence of the complementarity between dynamic sample weighting and multi-scale attention.

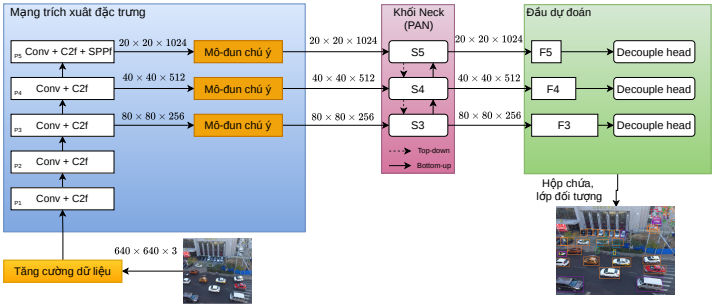


Figure 3.1. Architecture of the proposed LWMA-Net

Figure 3.1 illustrates the overall LWMA-Net architecture, in which feature extraction, multi-scale fusion, and attention modules are organized to prioritize small-target detection while maintaining a low parameter count. The architecture visually summarizes the principal design contribution presented in this chapter.

Chapter 4. Violence Detection in Video through Spatio-Temporal Characteristics

Chapter 4 presents a human-centric approach to violence detection. Person detection and tracking are used to determine the region of interest associated with each individual. Optical flow is subsequently computed within these regions and summarized using 25 statistical features describing motion magnitude and orientation over 15-frame windows. RGB images and appearance features are not directly provided to the classifier. Accordingly, the term spatio-temporal refers to representing motion within each person's spatial region and modeling the temporal evolution of that representation.

To prevent leakage caused by overlapping neighboring windows, the data are partitioned by group using a 70/10/20 training/validation/test ratio. For AIRTLab, the two videos of the same scene captured by different cameras are assigned to the same group. For Movies Fight, each video constitutes an independent group. All windows and tracklets from a group belong exclusively to one subset, and an automated audit confirms that no source overlap exists among the subsets. Eight classifiers and one ensemble are evaluated using three random seeds, 42, 1, and 2, and the results are reported as mean $\pm$ standard deviation at both window and video levels.

On AIRTLab with stride 1, the ensemble of 3D-CNN, CNN–LSTM, and Transformer obtains a macro-F1 score of $64.95 \pm 1.28\%$ and a video-level accuracy of $85.71 \pm 2.86\%$.

With stride 5, the corresponding values are $65.13 \pm 1.70\%$ and $80.48 \pm 2.97\%$. On Movies Fight, 3D-CNN achieves the highest macro-F1 score of $89.22 \pm 3.88\%$; XGBoost achieves the highest video-level accuracy of $98.33 \pm 1.44\%$; and the ensemble obtains the highest recall of $90.01 \pm 3.55\%$. These results indicate substantial domain dependence, and no single model is consistently superior across all metrics.

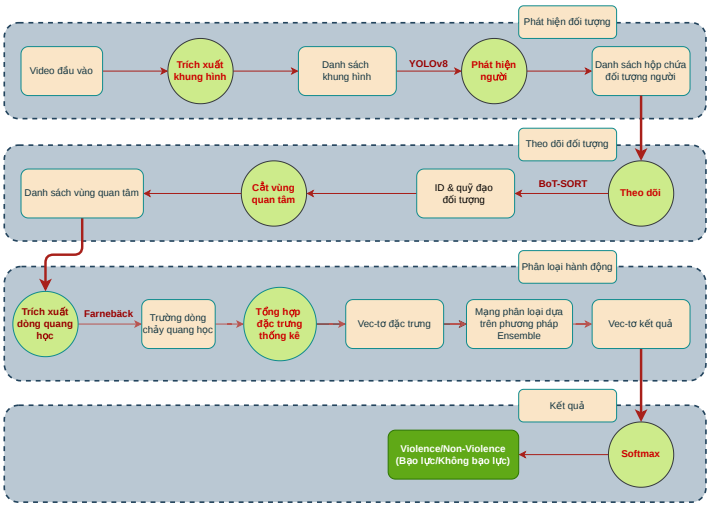


Figure 4.1. Overall architecture of the violence-detection system

Figure 4.1 presents the main processing stages, including person detection, object tracking, optical-flow extraction, statistical feature aggregation, and action classification. Computational analysis on an NVIDIA RTX 3080 10 GB GPU shows that all classifiers have a latency below 11 ms per window. Nevertheless, person detection, tracking,

and optical-flow extraction constitute the principal computational bottleneck and therefore require further optimization for real-time processing.

CONCLUSIONS AND FUTURE DIRECTIONS

1. Conclusions

The dissertation investigates and proposes methods for enhancing the efficiency of object and action recognition from surveillance cameras, with emphasis on three core challenges: imbalanced data, small-object detection, and violent-action recognition in surveillance images and videos. The principal results are summarized as follows.

First, the dissertation proposes DFL, a dynamic weighting strategy based on progressively increasing the focusing parameter γ during training. The same configuration is applied to classification, detection, and segmentation, demonstrating its integration potential across multiple computer-vision tasks. Experimental results show positive trends, particularly for difficult samples and small objects, while also indicating that DFL does not improve every metric uniformly.

Second, the dissertation develops LWMA-Net, a lightweight multi-scale attention model for small-object detection in UAV imagery. The model achieves 29.0% mAP and 47.4% mAP@50 on VisDrone2019. Integrating DFL increases these values to 29.5% mAP and 48.3% mAP@50 without changing inference complexity. This experiment provides direct evidence of the relationship and complementarity between the contributions of Chapters 2 and

3.

Third, the dissertation proposes a violence-detection framework based on motion representations within person regions. The classifier input consists of temporal sequences of statistical optical-flow features and does not include an independent RGB appearance branch. Group-based data partitioning eliminates source overlap among the subsets, while experiments on AIRTLab and Movies Fight using three random seeds provide a more cautious evaluation of generalization. The results also show that classification latency is low, whereas person detection, tracking, and optical-flow extraction remain the principal computational bottlenecks.

In summary, the dissertation accomplishes its stated research objectives and contributes solutions to three complementary aspects of object and action recognition from surveillance cameras. The results demonstrate the potential of the proposed methods, while deployment in operational systems still requires further validation using diverse surveillance data and under varying operating conditions.

The chapters are not organized as a strictly sequential pipeline in which the result of each chapter must be directly integrated into the next. Instead, they address complementary aspects of the overall problem. Chapter 2 proposes a general learning mechanism for imbalanced data, whereas Chapters 3 and 4 develop specialized solutions for different object- and action-recognition contexts. The methodological role of Chapter 2 is empirically demonstrated through the initial integration of DFL into LWMA-Net. Chapter 4 is

evaluated separately to isolate the effects of motion representation and temporal modeling.

2. Future Directions

Based on the achieved results and remaining limitations, the dissertation identifies the following future research directions.

Extending the integration of DFL and LWMA-Net into the video pipeline: The dissertation integrates DFL into LWMA-Net and observes improvements on VisDrone2019 without modifying the inference architecture. Future work should incorporate this object detector and imbalanced-learning mechanism into the video pipeline and evaluate their effects on tracklet quality and action recognition.

Extending evaluation to real-world surveillance data: AIRTLab and Movies Fight are both influenced by staged scenarios. The violence-detection method should therefore be evaluated on real-world surveillance datasets such as RWF-2000 or RLVS. Cross-dataset experiments, in which training and testing are performed on different domains, should also be conducted to investigate domain dependence.

Extending the methods to other object and action categories: The current models remain dependent on training-data quality, illumination, camera viewpoint, occlusion, scene diversity, and distribution shifts between training data and operational environments. Deployment in a new

environment may therefore require recalibration, reevaluation, and context-specific data. Future studies may extend the methods to unauthorized intrusion, abnormal gathering, falls, fire and explosion events, theft, and other hazardous behaviors. Such extensions should combine object detection, action recognition, and contextual modeling to improve applicability across diverse surveillance scenarios.

Completing statistical validation and reliability analysis: Chapter 4 is evaluated using three random seeds and reports mean values with standard deviations. For the computationally intensive experiments in Chapters 2 and 3, future work should conduct additional independent runs, report confidence intervals, and perform statistical significance testing, especially when differences between configurations are below one percentage point.

Investigating domain transfer and few-shot learning: Domain-adaptation and few-shot learning methods should be developed so that models can adapt rapidly to new environments and learn new abnormal-event categories from limited samples.

Improving explainability and decision support: Visualization should move beyond illustrating model-attention regions and support operators in understanding the reasons behind alerts. Attention maps, motion trajectories, object information, and concise semantic descriptions may be combined to provide visual evidence for individual system decisions.

Exploiting unlabeled data and auxiliary in-

formation: Large volumes of unlabeled surveillance video may be exploited through self-supervised learning, contrastive learning, or pretraining on target-domain data. Auxiliary information, including time, camera location, regions of interest, and scene context, may also improve system stability.

Optimizing system deployment and operation: For violence detection, optimization should prioritize GPU acceleration of person detection, tracking, and optical-flow extraction because these components account for most of the processing time. Future work should also investigate confidence-based alert mechanisms, post-deployment model updating, and workload evaluation for concurrent multi-camera streams.

Integration with intelligent surveillance infrastructure: In practical applications, recognition models must connect to camera-management, video-storage, event-retrieval, and real-time alert components. A modular system architecture should therefore be developed to support camera scalability, data-access control, and integration with existing security-management platforms.

LIST OF PUBLICATIONS RELATED TO THE DISSERTATION

[P1] **Nguyễn Dũng***, Hoàng Văn Dũng, Lê Văn Tường Lân (2024). *A Survey of Deep Learning-Based Object Detection Models*. Journal of Science and Technology, University of Sciences, vol. 23, no. 1, pp. 39–52 (Published).

[P2] **Dung Nguyen**, Van-Dung Hoang*, Van-Tuong-Lan Le (2025). *Evaluation of the effectiveness of modern object detection models on spatial image data*. Hue University Journal of Science. DOI: 10.26459/hueunijtt.v134i2B.7862 (Published).

[P3] **Nguyễn Dũng**, Hoàng Văn Dũng, Lê Văn Tường Lân* (2024). *Enhancing the DETR Object Detection Model by Integrating Linformer*. FAIR2024.
DOI: 10.15625/vap.2024.0210 (Published).

[P4] **Dung Nguyen**, Van-Dung Hoang*, and Van-Tuong-Lan Le (2024). *V-DETR: Pure Transformer for End-to-End Object Detection*. ACIIDS 2024, LNCS 14796, Springer (Scopus Q2). DOI: 10.1007/978-981-97-4985-0_10 (Published).

[P5] **Dung Nguyen**, Van-Dung Hoang*, Van-Tuong-Lan Le (2026). *Enhancing object detection efficiency with Transformers through multi-level feature integration*. Journal of Computer Science and Cybernetics. DOI: 10.15625/1813-9663/21114 (Published).

[P6] **Dung Nguyen**, Van-Dung Hoang*, Van-Tuong-Lan Le (2025). *Applying deep learning models for weapon detection in public environments through surveillance cameras*. Journal of Research and Development on ICT. DOI: 10.32913/mic-

ict-research.v2024.n2.1318 (Published).

[P7] **Dung Nguyen**, Van-Dung Hoang*, and Van-Tuong-Lan Le (2026). *Towards enhancing learning on imbalanced data: A novel adaptive weighting strategy*. Neurocomputing, vol. 673 (SCIE, Q1, IF 6.7). DOI: 10.1016/j.neucom.2026.132886 (Published).

[P8] **Dung Nguyen**, Van-Dung Hoang, and Van-Tuong-Lan Le* (2025). *A Lightweight Transformer Model For Real-Time Object Detection On Edge Devices*. ACIIDS 2025 (Scopus Q4). DOI: 10.1007/978-981-96-5884-8_16 (Published).

[P9] **Dung Nguyen**, Van-Dung Hoang*, and Van-Tuong-Lan Le (2026). *A Lightweight Multi-Scale Attention Model for Small Object Detection in UAV Imagery*. IEEE Access (SCIE, Q1, IF 4.2). DOI: 10.1109/ACCESS.2026.3656179 (Published).

[P10] **Dung Nguyen**, Van-Dung Hoang*, Van-Tuong-Lan Le (2026). *Tracklet-Driven Spatio-Temporal Representation Learning with Enhancing Optical Flow and Transformer Feature for Violence Detection*. Engineering Applications of Artificial Intelligence (Submitted).