ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN TỐT NGHIỆP

NGÀNH: CÔNG NGHỆ THÔNG TIN
CHUYÊN NGÀNH: KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO

ĐỀ TÀI:

Hệ thống phát hiện té ngã qua camera trên thiết bị biên ứng dụng kỹ thuật chắt lọc tri thức

Người hướng dẫn: **TS. NINH KHÁNH DUY**
Sinh viên thực hiện: **PHẠM QUANG NHỰT**
Số thẻ sinh viên: **102210351**
Lớp**: 21TCLC_NHAT1**

Đà Nẵng, 06/2025

TÓM TẮT

Tên đề tài: Hệ thống phát hiện té ngã qua camera trên thiết bị biên ứng dụng kỹ thuật chắt lọc tri thức
Sinh viên thực hiện: Phạm Quang Nhựt
Số thẻ SV: 102210351 Lớp: 21TCLC_NHAT1

Đề tài "Hệ thống phát hiện té ngã qua camera trên thiết bị biên ứng dụng kỹ thuật chắt lọc tri thức" tập trung vào việc phát triển một hệ thống phát hiện té ngã chính xác và hiệu quả, phù hợp triển khai trên các thiết bị biên có tài nguyên tính toán hạn chế. Té ngã là một trong những nguyên nhân gây chấn thương nghiêm trọng ở người cao tuổi, do đó việc phát hiện sớm có ý nghĩa quan trọng trong việc can thiệp và giảm thiểu hậu quả. Phương pháp được đề xuất sử dụng kỹ thuật truyền tri thức (knowledge distillation) nhằm chuyển giao kiến thức từ một mô hình lớn có hiệu suất cao (mô hình giáo viên) sang một mô hình nhỏ gọn hơn (mô hình trò), giúp giảm đáng kể độ phức tạp mô hình trong khi vẫn duy trì độ chính xác ở mức cao. Hệ thống được kiểm thử trên nhiều tập dữ liệu công khai với các chiến lược đánh giá chéo theo nền cảnh và theo người, cho thấy mô hình trò có thể tăng điểm F1 đến 7% và giảm kích thước xuống chỉ còn 1/200 so với mô hình ban đầu. Kết quả này chứng minh tính khả thi và tiềm năng ứng dụng của hệ thống trong các thiết bị giám sát thực tế.

ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Họ tên sinh viên: Phạm Quang Nhựt Số thẻ sinh viên: 102210351

Lớp: 21TCLC_NHAT1 Khoa: Công nghệ thông tin Ngành: Công nghệ thông tin (ngoại ngữ Nhật)

1. *Tên đề tài đồ án:* Hệ thống phát hiện té ngã qua camera trên thiết bị biên ứng dụng kỹ thuật chắt lọc tri thức
2. *Đề tài thuộc diện:* ☐ *Có ký kết thỏa thuận sở hữu trí tuệ đối với kết quả thực hiện*
3. *Các số liệu và dữ liệu ban đầu:*

   Không có
4. *Nội dung các phần thuyết minh và tính toán:*

   Mở đầu
   - **Chương 1**. Tổng quan đề tài.
   - **Chương 2**. Cơ sở lý thuyết & các nghiên cứu liên quan
   - **Chương 3**. Giải pháp đề xuất.
   - **Chương 4**. Kết quả thực nhiệm.

   Kết luận và hướng phát triển
5. *Các bản vẽ, đồ thị (ghi rõ các loại và kích thước bản vẽ):*

   Không Có
6. *Họ tên người hướng dẫn:* TS. Ninh Khánh Duy
7. *Ngày giao nhiệm vụ đồ án: …… /……/2025*
8. *Ngày hoàn thành đồ án: …… /……/2025*

Đà Nẵng, ngày tháng năm 2025

Trưởng Bộ môn……. **Người hướng dẫn**

LỜI CẢM ƠN

Trong hành trình xây dựng đồ án tốt nghiệp "Hệ thống phát hiện té ngã qua camera trên thiết bị biên ứng dụng kỹ thuật chắt lọc tri thức", em muốn bày tỏ lòng biết ơn sâu sắc đến những người đã đồng hành và hỗ trợ em, không chỉ về mặt kiến thức mà còn là nguồn động viên quan trọng.

Đầu tiên, em xin gửi lời cảm ơn chân thành đến thầy Ninh Khánh Duy, người đã đồng hành cùng em tại Trường Đại Học Bách Khoa Đà Nẵng, với sự hướng dẫn, nhận xét và hỗ trợ quý báu của thầy.

Ngoài ra, em cũng muốn bày tỏ lòng biết ơn đặc biệt đến các thầy cô trong Khoa Công nghệ thông tin, những người đã chia sẻ ý kiến quý báu và hỗ trợ em trong quá trình nghiên cứu và triển khai dự án.

Với thời gian và kiến thức có hạn, em nhận thức rằng đồ án không tránh khỏi những thiếu sót. Mong rằng, sự đóng góp ý kiến của các thầy cô trong nhà trường và bạn bè sẽ giúp em rút kinh nghiệm và hoàn thiện công trình của mình.

Em xin chân thành cảm ơn!

Phạm Quang Nhựt

CAM ĐOAN

1. Nội dung trong đồ án này là do em thực hiện dưới sự hướng dẫn trực tiếp của TS. Ninh Khánh Duy. Các kết quả trong đồ án đều được thực hiện trong quá trình tôi đang học tập tại Khoa Công Nghệ Thông Tin, trường Đại học Bách Khoa – Đại học Đà Nẵng.

2. Các tham khảo dùng trong đồ án đều được trích dẫn rõ ràng tên tác giả, tên công trình, thời gian, địa điểm công bố, được kèm với đường dẫn hợp lệ.

3. Nếu có những sao chép không hợp lệ, vi phạm quy chế đào tạo, em xin chịu hoàn toàn trách nhiệm.

Sinh viên thực hiện

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

DANH SÁCH CÁC KÝ HIỆU, CHỮ VIẾT TẮT

Từ viết tắt	Diễn giải
KD	Knowledge Distillation – Chắt lọc tri thức
ST-GCN	Spatial Temporal Graph Convolutional Networks - Mạng Nơ-ron Tích Chập Đồ Thị Không-Gian Thời-Gian
FD	Fall Detection - Phát hiện té ngã
CV	Computer Vision - Thị giác máy tính
HAR	Human Activity Recognition - Nhận diện hành động con người
MSE	Mean Squared Error – Sai số bình phương trung bình
HPE	Human Pose Estimation – Ước lượng tư thế cơ thể người
FPS	Frames Per Second – Số khung hình mỗi giây
GCN	Graph Convolutional Network – Mạng tích chập đồ thị
CNN	Convolutional Neural Network – Mạng nơ-ron tích chập
RNN	Recurrent Neural Network – Mạng nơ-ron hồi tiếp
LSTM	Long Short-Term Memory – Bộ nhớ ngắn dài hạn
ONNX	Open Neural Network Exchange – Định dạng chuyển đổi mô hình học sâu
IoT	Internet of Things – Internet vạn vật
KL	Kullback-Leibler Divergence – Độ lệch phân phối

MỞ ĐẦU

1. Tổng quan về đề tài

Té ngã là một trong những nguyên nhân hàng đầu gây chấn thương nghiêm trọng ở người cao tuổi, không chỉ ảnh hưởng đến sức khỏe mà còn làm giảm chất lượng cuộc sống, thậm chí có thể dẫn đến tử vong nếu không được phát hiện và xử lý kịp thời. Các hệ thống phát hiện té ngã từ xa hiện nay vẫn còn nhiều hạn chế, đặc biệt là trong môi trường thực tế, nơi điều kiện ánh sáng, góc quay camera và khả năng tính toán của thiết bị không phải lúc nào cũng lý tưởng.Do đó, việc nghiên cứu và áp dụng kỹ thuật chắt lọc tri thức (knowledge distillation) nhằm tối ưu hóa mô hình phát hiện té ngã trở nên vô cùng quan trọng. Bằng cách tạo ra một mô hình nhẹ hơn nhưng vẫn đảm bảo hiệu suất cao, hệ thống có thể hoạt động hiệu quả trên các thiết bị biên (edge devices), giúp phát hiện té ngã nhanh chóng và đáng tin cậy. Nghiên cứu này không chỉ mang ý nghĩa khoa học mà còn có tiềm năng ứng dụng rộng rãi trong thực tế, hỗ trợ chăm sóc và bảo vệ người cao tuổi một cách hiệu quả hơn.

2. Mục đích và ý nghĩa của đề tài

2.1. Mục đích

- Xây dựng và triển khai thành công một mô hình nhận diện té ngã dựa trên dữ liệu khung xương, sử dụng ST-GCN kết hợp với KD để tối ưu hóa hiệu suất.
- So sánh hiệu năng và chất lượng giữa các phương pháp khác nhau nhằm tìm ra hướng tối ưu cho ứng dụng thực tế.

2.2. Ý nghĩa và phạm vi nghiên cứu

- Ý nghĩa**:** Giúp người dùng nhận diện té ngã dễ dàng hơn: Hệ thống hỗ trợ phát hiện và cảnh báo khi người dùng gặp sự cố té ngã, đặc biệt là người cao tuổi hoặc người có nguy cơ cao, giúp giảm thiểu chấn thương nghiêm trọng.
- Phạm vi**:** Đề tài tập trung phát hiện té ngã từ dữ liệu khung xương được trích xuất từ video, không sử dụng ảnh RGB hay cảm biến khác. Dữ liệu thu thập đa dạng về góc nhìn và đối tượng, được đánh giá bằng kiểm định chéo theo người và góc nhìn. Mô hình hướng đến triển khai trên thiết bị biên để giám sát té ngã trong môi trường thực tế. Thời gian thực hiện khoảng 2,5 tháng.

3. Phương pháp thực hiện

a. Thu thập và tiền xử lý dữ liệu

- Sử dụng các bộ dữ liệu chuyên biệt về té ngã
- Trích xuất dữ liệu khung xương từ video bằng MediaPipe Pose hoặc OpenPose.
- Chuẩn hóa dữ liệu bằng cách loại bỏ nhiễu, nội suy khung hình bị mất và căn chỉnh dữ liệu khung xương.
- Chia tập dữ liệu thành Train, Validation, Test theo hai phương pháp:
    - Cross-view: Kiểm tra mô hình trên góc nhìn khác với dữ liệu huấn luyện.
    - Cross-person: Kiểm tra mô hình trên những người không có trong tập huấn luyện.

b. Xây dựng mô hình nhận diện té ngã

- Sử dụng mô hình Spatial-Temporal Graph Convolutional Network (ST-GCN) để học đặc trưng chuyển động từ dữ liệu khung xương.
- Áp dụng chắt lọc tri thức để tối ưu hóa mô hình bằng cách truyền tải thông tin từ một mô hình lớn (teacher model) sang mô hình nhỏ hơn (student model).
- Thử nghiệm các phương pháp giảm số lượng tham số của mô hình.

c. Huấn luyện và kiểm thử

- Huấn luyện mô hình trên tập dữ liệu chuẩn bị sẵn, sử dụng thuật toán tối ưu AdamW.
- Sử dụng Cross-Entropy Loss để đo lường độ chính xác của mô hình.
- Đánh giá hiệu suất mô hình dựa trên các chỉ số:
    - Accuracy, Precision, Recall, F1-score trên các tập kiểm thử.
    - Độ trễ và hiệu suất trên thiết bị biên để đảm bảo mô hình có thể chạy trong thời gian thực.

4. Cấu trúc của đồ án

Mở đầu: trình bày về tổng quan đề tài, mục đích, ý nghĩa đề tài, phương pháp nghiên cứu, cấu trúc của đồ án tốt nghiệp.

Chương 1: trình bày các nội dung cơ sở lý thuyết chính liên quan đến các giải pháp triển khai trong đồ án.

Chương 2: trình bày về việc cơ sở lý thuyết, các nghiên cứu phương pháp trước đây và điểm mạnh, điểm yếu mà phương pháp đề xuất cần khắc phục

Chương 3: Mô tả kiến trúc hệ thống đề xuất, quy trình thu thập và xử lý dữ liệu, cũng như phương pháp đánh giá mô hình.

Chương 4: Trình bày kết quả thực nghiệm trên nhiều bộ dữ liệu, hiệu quả của mô hình sau khi áp dụng chắt lọc tri thức và khả năng triển khai thực tế trên thiết bị biên.

Kết luận và hướng phát triển: Trình bày về kết quả và hướng phát triển.

Tài liệu tham khảo

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1. Bài toán

1.1.1. Giới thiệu bài toán

Té ngã là một hành động đột ngột, không chủ ý và thường gây ra hậu quả nghiêm trọng như chấn thương, gãy xương hoặc tổn thương nội tạng – đặc biệt nguy hiểm đối với người cao tuổi hoặc những người có vấn đề về vận động. Theo Tổ chức Y tế Thế giới (WHO), té ngã hiện đang là nguyên nhân gây tử vong do chấn thương không chủ ý đứng thứ hai trên toàn cầu, với khoảng 684.000 ca tử vong mỗi năm, trong đó hơn 80% xảy ra tại các quốc gia có thu nhập thấp và trung bình [1]. Ngoài ra, ước tính mỗi năm có khoảng 37,3 triệu trường hợp té ngã nghiêm trọng cần can thiệp y tế, dẫn đến tổng cộng hơn 38 triệu năm sống điều chỉnh theo mức độ tàn tật (DALYs) bị mất đi.

Tại Việt Nam, số lượng ca té ngã có xu hướng gia tăng theo độ tuổi. Cụ thể, tỷ lệ té ngã đạt đỉnh 32,7% ở nhóm người từ 80 tuổi trở lên, cho thấy người cao tuổi, đặc biệt là trên 70 tuổi, là đối tượng có nguy cơ cao nhất [2]. Nếu không được phát hiện và xử lý kịp thời, các biến chứng do té ngã có thể kéo dài đến 12 tháng, gây ảnh hưởng nghiêm trọng đến sức khỏe, tinh thần và chất lượng cuộc sống của bệnh nhân [3].

Té ngã thường xảy ra một cách bất ngờ trong các hoạt động sinh hoạt hằng ngày, nên rất khó dự đoán trước. Do đó, bài toán phát hiện té ngã tự động ngày càng thu hút sự quan tâm mạnh mẽ từ cộng đồng nghiên cứu nhờ tiềm năng ứng dụng thực tiễn trong các hệ thống chăm sóc sức khỏe thông minh, đặc biệt là giám sát người cao tuổi sống một mình hoặc bệnh nhân phục hồi chức năng có nguy cơ té ngã cao.

Tuy nhiên, việc phát hiện té ngã từ video vẫn đối mặt với nhiều thách thức. Một mặt, dữ liệu về té ngã thường hiếm gặp, mang tính bất thường và khó tái hiện chính xác trong điều kiện thực tế, dẫn đến hiện tượng mất cân bằng dữ liệu trong quá trình huấn luyện. Mặt khác, nhu cầu triển khai hệ thống trên các thiết bị biên (edge devices) – vốn có giới hạn về bộ nhớ và khả năng xử lý – đòi hỏi các mô hình phải được tối ưu hóa mạnh mẽ về mặt hiệu năng.

Bài toán đặt ra là phát hiện hành vi té ngã từ một chuỗi hình ảnh đầu vào. Cụ thể:

- Đầu vào: Một sequence gồm N khung hình liên tiếp (thường là 32 frame), được trích xuất từ video. Số lượng khung hình N có thể điều chỉnh linh hoạt tùy theo thiết lập thí nghiệm. Sau mỗi chuỗi, cửa sổ trượt sẽ di chuyển một bước (tức là dịch sang frame tiếp theo) để tạo ra chuỗi kế tiếp nhằm bao phủ toàn bộ video.
- Đầu ra: Nhãn đầu ra của mỗi chuỗi là một trong hai lớp: 1: Có xảy ra té ngã trong khoảng thời gian tương ứng với sequence. 0: Không có té ngã trong sequence đó.

Mục tiêu của bài toán là xây dựng một mô hình học máy có khả năng nhận diện chính xác các hành vi té ngã từ chuỗi hình ảnh theo thời gian, đảm bảo hiệu suất cao cả về độ chính xác và tốc độ suy luận, đặc biệt là khi triển khai trên thiết bị có tài nguyên giới hạn như thiết bị di động hoặc thiết bị biên.

1.1.2. Tại sao phải phát hiện té ngã

Việc phát hiện té ngã có vai trò quan trọng trong nhiều lĩnh vực, đặc biệt là chăm sóc sức khỏe và đảm bảo an toàn cho con người. Té ngã là một trong những nguyên nhân hàng đầu gây ra thương tích nghiêm trọng, đặc biệt ở người cao tuổi, người mắc bệnh huyết áp, tim mạch hoặc các rối loạn vận động như Parkinson hay thoái hóa khớp. Những đối tượng này thường có nguy cơ cao bị té ngã khi sinh hoạt một mình, và nếu không được phát hiện và hỗ trợ kịp thời, hậu quả có thể rất nghiêm trọng, thậm chí dẫn đến tử vong. Do đó, việc xây dựng một hệ thống phát hiện té ngã tự động và cảnh báo nhanh chóng là hết sức cần thiết nhằm hỗ trợ giám sát và can thiệp kịp thời.

Trong các hệ thống an ninh – an toàn, phát hiện hành vi té ngã cũng góp phần cảnh báo sớm những tình huống nguy hiểm hoặc bất thường tại các khu vực cần được giám sát nghiêm ngặt như viện dưỡng lão, bệnh viện hay các khu vực cách ly. Các hệ thống này thường kết hợp cảm biến và thị giác máy tính để nhận diện hành vi bất thường và phát ra cảnh báo nếu có người gặp sự cố.

Để giải quyết bài toán này, nhiều hệ thống phát hiện té ngã đã được đề xuất, bao gồm thiết bị đeo (wearable) và không đeo (non-wearable) [4, 5, 6]. Các hệ thống wearable mang lại độ chính xác cao hơn nhờ được gắn trực tiếp lên cơ thể người dùng, tuy nhiên lại gây khó chịu trong quá trình sử dụng lâu dài, dẫn đến dữ liệu thu thập không ổn định và ảnh hưởng đến hiệu quả mô hình [7]. Ngoài ra, giới hạn về thời lượng pin khiến các thiết bị này khó đáp ứng được yêu cầu giám sát liên tục [8, 9].

Trong khi đó, các hệ thống non-wearable như cảm biến sàn (ví dụ Ground Reaction Force - GRF) [4] hoặc giải pháp dựa trên thị giác máy tính [5] đang được ứng dụng rộng rãi nhờ khả năng thu thập thông tin hình ảnh phong phú và không cần tiếp xúc vật lý [10]. Phương pháp này cho phép giám sát liên tục và ít xâm lấn, tuy nhiên việc xử lý video thời gian thực lại đòi hỏi phần cứng mạnh và khả năng thích ứng với nhiều môi trường khác nhau [11, 12].

1.2. Các nghiên cứu có liên quan

Các phương pháp phát hiện té ngã thời gian thực hiện nay được chia thành ba nhóm chính: phương pháp dựa trên ngưỡng, học máy truyền thống và mô hình học sâu nhẹ (lightweight deep learning). Phương pháp ngưỡng sử dụng các quy tắc định sẵn dựa trên dữ liệu RGB, độ sâu, hoặc góc nghiêng cơ thể và có thể đạt độ chính xác lên đến 0.97 [13, 14, 15, 16], nhưng lại dễ nhầm lẫn với các hành vi như nằm xuống hoặc cúi người. Các mô hình học máy như SVM và KNN khi áp dụng lên các điểm đặc trưng cơ thể hoạt động tốt trên khung hình tĩnh nhưng không hiệu quả với chuỗi hành động liên tục, dẫn đến việc phân loại sai [17]. Cuối cùng, các phương pháp học sâu như CNN và LSTM cải thiện độ chính xác bằng cách học các đặc trưng không gian và thời gian, tuy nhiên cần được tối ưu để đảm bảo hiệu năng thời gian thực [18, 19, 20].

1.2.1. Theo công nghệ cảm biến

Trong lĩnh vực nhận dạng hoạt động và đặc biệt là phát hiện té ngã, cảm biến đóng vai trò then chốt. Chúng giúp theo dõi chuyển động cơ thể, môi trường và các thông số liên quan từ xa, thường sử dụng các giao thức truyền dữ liệu không dây như Wi-Fi, Bluetooth, cho phép thu thập dữ liệu liên tục và không gây cản trở cho người dùng.

Nhờ sự phát triển nhanh chóng của công nghệ, cảm biến ngày càng nhỏ gọn, tiết kiệm năng lượng, bền bỉ và ít bị ảnh hưởng bởi điều kiện môi trường. Điều

này giúp các cảm biến dễ dàng được tích hợp vào các thiết bị đeo như đồng hồ thông minh, điện thoại, vòng tay,... Các thiết bị này cho phép theo dõi hoạt động của người dùng trong thời gian dài mà không gây khó chịu, đồng thời không bị giới hạn bởi không gian cố định như camera lắp trong phòng.

Một số nghiên cứu đã chứng minh hiệu quả của cảm biến đeo trong nhận dạng hoạt động con người. Ví dụ, [21] sử dụng hai thiết bị Wii Remote đeo tại tay và thắt lưng để phân biệt 14 hoạt động thường ngày với độ chính xác và độ bao phủ trên 90%

Trong các cảm biến đeo, cảm biến quán tính (IMU) bao gồm gia tốc kế, con quay hồi chuyển và từ kế là phổ biến nhất. Chúng thường được tích hợp trong các thiết bị hằng ngày và có thể nhận dạng các hoạt động như đi, chạy, ngồi, nhảy, hoặc phát hiện tư thế thay đổi bất thường. Việc kết hợp gia tốc kế và con quay hồi chuyển giúp tăng độ chính xác trong việc phân biệt té ngã với các hoạt động thông thường như ngồi xuống nhanh hoặc bước hụt cầu thang. Một số nghiên cứu còn kết hợp cảm biến áp suất với gia tốc kế để tăng khả năng phát hiện chính xác hành vi té ngã [22].

Ngoài ra, các cảm biến như:

- Từ kế: xác định hướng chuyển động, hỗ trợ phân biệt người dùng đang nằm, ngồi hay đứng.
- Cảm biến điện cơ (EMG): phát hiện hoạt động cơ bắp, giúp nhận diện ngã với độ chính xác cao nhưng ít được sử dụng do kích thước lớn, khó đeo trong thực tế.
- Cảm biến rung và cảm biến điện từ gắn trên sàn nhà: có thể phát hiện ngã với độ chính xác cao (lên đến 100% [23]), nhưng chi phí thiết lập hệ thống cao, khó triển khai diện rộng.
- Cảm biến hình ảnh (ví dụ: SenseCam, Kinect): ghi lại hoạt động của người dùng, nhưng đặt ra các thách thức lớn về quyền riêng tư, lưu trữ và xử lý dữ liệu.

Hiện nay, nhiều thiết bị thương mại đã tích hợp cảm biến để hỗ trợ chăm sóc sức khỏe. Chẳng hạn, Apple Watch có thể phát hiện ngã và tự động gửi cảnh báo, tuy nhiên giá thành còn cao và phụ thuộc vào hệ sinh thái thiết bị Apple. Một số ứng dụng như Moves kết hợp dữ liệu từ gia tốc kế và GPS để theo dõi hành vi di chuyển của người dùng.

Tuy nhiên, các thiết bị này vẫn còn nhiều hạn chế:

- Đa phần chỉ hỗ trợ theo dõi hoạt động thể chất đơn giản (bước đi, lượng calorie).
- Dữ liệu thu thập còn nhiễu do thay đổi vị trí đeo, môi trường, hoặc lỗi đo từ pin.
- Các hệ thống phát hiện té ngã chuyên biệt vẫn chưa phổ biến tại Việt Nam.

Do đó, vẫn cần tiếp tục nghiên cứu những phương pháp phát hiện té ngã hiệu quả hơn, tối ưu cho các thiết bị đeo hoặc triển khai trên thiết bị biên với chi phí thấp, tính khả thi cao trong thực tế.

1.2.2. Theo công nghệ thị giác máy tính

Phát hiện té ngã (fall detection) là một lĩnh vực đã được nghiên cứu sâu rộng trong nhiều năm qua, với nhiều phương pháp tiếp cận khác nhau, trong đó

các phương pháp dựa trên thị giác máy tính (vision-based) sử dụng dữ liệu hình ảnh RGB, ảnh chiều sâu (depth) và hồng ngoại (infrared - IR) đóng vai trò then chốt. Trong ba loại dữ liệu này, ảnh RGB vẫn là lựa chọn phổ biến nhất nhờ vào tính sẵn có rộng rãi, chi phí thấp và khả năng tích hợp dễ dàng với các thiết bị ghi hình thông thường như camera giám sát [24].

Những năm gần đây, với sự phát triển vượt bậc của học sâu (deep learning), nhiều nghiên cứu đã khai thác các mô hình học sâu để phân tích đặc trưng không gian - thời gian (spatio-temporal features) trong chuỗi hình ảnh hoặc video, nhằm nhận diện chính xác các tình huống té ngã [25, 26]. Nhờ đó, khả năng nhận diện hành vi bất thường hoặc nguy hiểm như ngã đã đạt được độ chính xác cao hơn, kể cả trong môi trường phức tạp và có nhiều biến động.

Bên cạnh đó, một hướng tiếp cận khác đang ngày càng phổ biến trong nhận dạng hoạt động con người (Human Action Recognition - HAR) và phát hiện té ngã là phương pháp dựa trên bộ xương (skeleton-based). Những phương pháp này có nhiều ưu điểm nổi bật như khả năng biểu diễn đặc trưng hiệu quả, dễ diễn giải (interpretability), và ít phụ thuộc vào bối cảnh hình ảnh [24]. Chúng thường sử dụng các thuật toán ước lượng tư thế cơ thể người (pose estimation) như OpenPose [27] hoặc AlphaPose [28] để trích xuất các điểm mốc trên cơ thể từ hình ảnh hoặc video. Nhờ đó, hệ thống có thể phân tích cả thông tin không gian (vị trí các khớp) và thông tin thời gian (diễn tiến của chuyển động).

Trong các hệ thống nhận dạng hoạt động dựa trên bộ xương, nhiều mô hình học sâu đã được ứng dụng để khai thác mối quan hệ theo chuỗi thời gian hoặc mối quan hệ giữa các khớp, như Mạng bộ nhớ dài ngắn hạn (LSTM) và Mạng nơ-ron đồ thị (GCN) [29, 30]. Những mô hình này cho phép hệ thống hiểu được mối liên kết phức tạp trong chuyển động của con người theo thời gian, từ đó tăng cường khả năng phân biệt giữa các hoạt động bình thường và các hành vi nguy hiểm như té ngã.

Để hỗ trợ khả năng triển khai thực tế, đặc biệt là trên các thiết bị biên (edge devices), nhiều nghiên cứu đã tập trung vào việc giảm thời gian xử lý và độ phức tạp của mô hình, nhằm đảm bảo hiệu suất thời gian thực. Ví dụ, Ramirez et al. [17] đã sử dụng các mô hình học máy truyền thống như SVM, KNN, Random Forest (RF) và Multi-Layer Perceptron (MLP) ở tầng phân loại cuối cùng để giảm tải tính toán. Trong khi đó, Noor et al. [31] đã áp dụng kỹ thuật Post-Training Quantization để giảm kích thước và độ phức tạp của mô hình, giúp phù hợp hơn với thiết bị có tài nguyên hạn chế. Ngoài ra, nghiên cứu của Chang et al. [32] tập trung tối ưu hóa mô hình bằng cách giảm số lượng tham số, nhằm tăng độ tương thích với phần cứng nhúng.

Tuy nhiên, phần lớn các hệ thống phát hiện té ngã hiện nay chưa khai thác hiệu quả các kỹ thuật tối ưu hóa hiện đại trong quá trình huấn luyện, nhằm tìm được sự cân bằng giữa giảm độ phức tạp mô hình và duy trì độ chính xác cao. Chính vì vậy, nghiên cứu này đề xuất ứng dụng phương pháp chắt lọc tri thức (Knowledge Distillation - KD) để cải thiện khả năng phát hiện té ngã. Phương pháp này cho phép mô hình nhẹ hơn (student model) học lại kiến thức từ một mô hình lớn và mạnh hơn (teacher model), giúp duy trì hiệu quả cao trong khi vẫn đảm bảo khả năng chạy trên các thiết bị tài nguyên thấp.

1.3. Tổng quan về thiết bị biên

1.3.1. Khái niệm về điện toán biên

Trong bối cảnh nhu cầu xử lý dữ liệu thời gian thực ngày càng tăng, Edge

Computing (điện toán biên) nổi lên như một giải pháp hiệu quả nhằm giảm độ trễ, tiết kiệm băng thông và nâng cao khả năng phản hồi cục bộ của hệ thống. Thay vì gửi toàn bộ dữ liệu lên đám mây để xử lý, Edge Computing cho phép xử lý trực tiếp tại thiết bị gần nguồn phát sinh dữ liệu – còn gọi là thiết bị biên (edge device). Điều này đặc biệt quan trọng đối với các hệ thống phát hiện té ngã trong môi trường thực tế, nơi yêu cầu tốc độ xử lý nhanh và liên tục.

Trong hệ sinh thái điện toán biên, các thiết bị biên như camera thông minh, cảm biến chuyển động, hoặc thiết bị IoT gắn trên cơ thể sẽ thu thập dữ liệu và xử lý tại chỗ hoặc truyền về máy chủ biên (edge server) để phân tích thêm. Mô hình này giúp giảm tải cho hệ thống đám mây, đồng thời tăng tính bảo mật và độ tin cậy, ngay cả khi có sự cố mạng.

Lợi ích chính của Edge Computing trong bài toán phát hiện té ngã:

- Độ trễ thấp: Phát hiện và cảnh báo nhanh khi có té ngã nhờ xử lý ngay tại biên.
- Không phụ thuộc hoàn toàn vào kết nối Internet: Giảm thiểu rủi ro mất kết nối gây gián đoạn hệ thống.
- Tiết kiệm băng thông và tài nguyên đám mây: Dữ liệu chỉ được gửi lên đám mây khi thực sự cần thiết.
- Tăng cường bảo mật: Dữ liệu nhạy cảm như video giám sát có thể được xử lý cục bộ mà không cần gửi đi xa.

1.3.2. Thiết bị biên trong học sâu

Học sâu gần đây đã đạt được nhiều thành công vượt trội trong lĩnh vực máy học, đặc biệt trong các ứng dụng như thị giác máy tính, xử lý ngôn ngữ tự nhiên và phân tích dữ liệu lớn. Ví dụ, các phương pháp học sâu liên tục vượt trội so với các phương pháp truyền thống trong các bài toán nhận dạng và phát hiện đối tượng tại cuộc thi ISLVRC từ năm 2012 [34]. Tuy nhiên, độ chính xác cao của học sâu đi kèm với yêu cầu rất lớn về tính toán và bộ nhớ, cả trong giai đoạn huấn luyện và suy luận.

Việc huấn luyện một mô hình học sâu rất tốn kém về mặt tính toán và dung lượng do phải tinh chỉnh hàng triệu tham số qua nhiều giai đoạn. Trong khi đó, suy luận cũng tiêu tốn nhiều tài nguyên do đầu vào thường có độ phân giải cao và yêu cầu thực hiện hàng triệu phép tính. Đặc trưng nổi bật của học sâu là đạt độ chính xác cao nhưng đồng thời tiêu tốn tài nguyên lớn.

Một giải pháp phổ biến để đáp ứng yêu cầu tính toán của học sâu là sử dụng điện toán đám mây. Tuy nhiên, để sử dụng được tài nguyên này, dữ liệu từ các thiết bị đầu cuối nằm ở biên mạng (chẳng hạn như smartphone hoặc cảm biến IoT) cần được truyền tới một trung tâm xử lý tập trung trên đám mây. Cách tiếp cận này phát sinh một số thách thức:

- Độ trễ: Nhiều ứng dụng yêu cầu suy luận thời gian thực. Ví dụ, các khung hình từ camera trên xe tự hành cần được xử lý tức thời để phát hiện và tránh vật cản, hoặc các ứng dụng trợ lý giọng nói cần phản hồi nhanh chóng. Tuy nhiên, việc truyền dữ liệu lên đám mây để xử lý có thể gây thêm độ trễ hàng trăm mili-giây do hàng đợi và truyền dẫn, dẫn đến không đáp ứng được yêu cầu thời gian thực nghiêm ngặt. Thực nghiệm cho thấy việc gửi một khung hình camera lên dịch vụ Amazon Web Services để xử lý thị giác máy tính mất hơn 200 ms tính từ đầu đến cuối [35].
- Khả năng mở rộng: Việc gửi dữ liệu từ thiết bị đầu cuối lên đám mây tạo ra vấn đề về khả năng mở rộng khi số lượng thiết bị tăng nhanh. Điều này

có thể gây tắc nghẽn mạng và sử dụng không hiệu quả tài nguyên nếu như không phải tất cả dữ liệu đều cần được xử lý bởi mô hình học sâu. Điều này đặc biệt nghiêm trọng với các nguồn dữ liệu yêu cầu băng thông cao như luồng video.

- Riêng tư: Việc gửi dữ liệu lên đám mây làm dấy lên các lo ngại về quyền riêng tư của người dùng, đặc biệt khi dữ liệu chứa thông tin nhạy cảm như khuôn mặt hay giọng nói. Ví dụ, việc triển khai các camera và cảm biến tại thành phố thông minh ở New York đã gây ra nhiều lo ngại từ các tổ chức bảo vệ quyền riêng tư [36].

Điện toán biên (Edge Computing) được xem là một giải pháp khả thi để giải quyết các thách thức nêu trên. Trong mô hình này, các tài nguyên tính toán được triển khai gần với các thiết bị đầu cuối [37]. Ví dụ, một nút tính toán biên có thể được đặt cùng với trạm gốc di động, cổng IoT hoặc tại một mạng nội bộ như trường đại học. Hiện nay, các doanh nghiệp lớn đã bắt đầu triển khai điện toán biên – ví dụ, một nhà cung cấp dịch vụ di động lớn tại Mỹ và một chuỗi cửa hàng thức ăn nhanh quốc gia đã triển khai hệ thống tính toán biên riêng [38], [39].

- Để giải quyết vấn đề độ trễ, điện toán biên cho phép xử lý dữ liệu gần nguồn phát sinh, nhờ đó giảm thời gian truyền và hỗ trợ các dịch vụ thời gian thực.
- Với khả năng mở rộng, điện toán biên tạo nên một kiến trúc phân cấp giữa thiết bị đầu cuối, nút biên và trung tâm dữ liệu đám mây, từ đó giảm tải mạng trung tâm và tăng hiệu quả hệ thống.
- Đối với riêng tư, việc xử lý dữ liệu ngay tại thiết bị hoặc máy chủ biên đáng tin cậy giúp tránh việc truyền tải qua Internet công cộng và giảm rủi ro bị tấn công hay rò rỉ dữ liệu.

Tuy nhiên, để hiện thực hóa học sâu trên thiết bị biên, vẫn còn nhiều thách thức kỹ thuật cần giải quyết. Một trong những thách thức chính là việc đáp ứng yêu cầu tài nguyên lớn của học sâu trên các thiết bị biên có khả năng tính toán hạn chế. Các thiết bị này rất đa dạng, từ máy chủ biên có GPU đến smartphone với bộ xử lý di động, thậm chí là các thiết bị tối giản như Raspberry Pi. Thách thức thứ hai là làm thế nào để các thiết bị biên phối hợp với nhau và với đám mây trong điều kiện tài nguyên và mạng không đồng nhất, nhằm đảm bảo hiệu suất tối ưu ở cấp độ ứng dụng. Cuối cùng, dù điện toán biên giúp cải thiện vấn đề riêng tư, nhưng một số dữ liệu vẫn cần trao đổi giữa các thiết bị hoặc với đám mây, gây ra rủi ro bảo mật.

Các nhà nghiên cứu đã đề xuất nhiều cách tiếp cận khác nhau để giải quyết các thách thức này, từ thiết kế phần cứng, kiến trúc hệ thống đến mô hình lý thuyết và phân tích. Mục tiêu của bài viết gốc là khảo sát các nghiên cứu nằm tại giao điểm giữa hai xu hướng lớn: học sâu và điện toán biên, đặc biệt tập trung vào khía cạnh phần mềm và các thách thức đặc thù trong lĩnh vực này. Mặc dù đã có nhiều khảo sát chuyên sâu riêng biệt về học sâu [40] và điện toán biên [41], [42], bài viết này tập trung vào phần giao nhau giữa hai lĩnh vực.

Học sâu trên thiết bị biên có nhiều điểm tương đồng nhưng cũng có sự khác biệt với các chủ đề nghiên cứu truyền thống khác. So với điện toán đám mây (ví dụ như mô hình “machine learning as a service”), điện toán biên mang lại lợi ích về độ trễ thấp và khả năng định vị theo không gian – những lợi thế đã được tận dụng trong nhiều nghiên cứu [43]. Một số công trình còn kết hợp cả điện

toán biên và điện toán đám mây, tạo nên các kiến trúc lai edge-cloud [44]. So với các phương pháp học máy truyền thống (ngoài học sâu), học sâu đòi hỏi tài nguyên lớn hơn, nhưng cấu trúc nội tại của học sâu có thể được khai thác để giảm tải tính toán [45, 46]. Ngoài ra, điện toán biên còn có những thách thức đặc thù do việc chia sẻ tài nguyên tính toán và truyền thông giữa nhiều thiết bị.

1.3.3. Một số nghiên cứu tối ưu trên thiết bị biên

Nhiều nghiên cứu đã tập trung vào việc giảm độ trễ của deep learning khi triển khai trên thiết bị tài nguyên hạn chế (xem Fig. 1), từ đó cải thiện hiệu năng cho toàn hệ sinh thái edge. Một hướng tiếp cận phổ biến là thiết kế mô hình với ít tham số hơn nhằm giảm bộ nhớ và thời gian thực thi mà vẫn giữ được độ chính xác cao, ví dụ như MobileNets [21], SSD [69], YOLO [18] và SqueezeNet [70], các mô hình này có sẵn trên TensorFlow [24] và Caffe [28]. Một hướng khác là nén mô hình, gồm lượng tử hóa tham số (biến số thực thành số nguyên độ dài bit thấp để giảm phép toán), cắt tỉa tham số (loại các trọng số ít quan trọng) [71], [73], và rút gọn kiến thức (huấn luyện mạng nhỏ bắt chước mạng lớn) [75]. DeepIoT [34] đề xuất phương pháp pruning cho thiết bị IoT, Lai và Suda [72] cung cấp CMSIS-NN cho ARM Cortex-M, Han et al. [73] kết hợp pruning và quantization cho RNN giúp tăng tốc 10× và 2×, Bhattacharya và Lane [74] sparsify mạng fully connected cho thiết bị đeo. Adadeep [77] tự động chọn kỹ thuật nén phù hợp tài nguyên, DeepMon [78] kết hợp quantization và cache tầng trung gian trên GPU để tránh tính toán lại giữa các khung hình. Về phần cứng, các hãng đã phát triển chip chuyên dụng như Google TPU [79], ShiDianNao [80] (nhắm đến truy cập bộ nhớ hiệu quả cho thiết bị nhúng), FPGA [82] giúp xử lý nhanh và tiết kiệm điện hơn so với CPU/GPU truyền thống. Các công cụ hỗ trợ đi kèm gồm Intel OpenVINO [83], [84], Nvidia EGX [85], Qualcomm Neural Processing SDK [86], RSTensorFlow [87]. Lane et al. [88] còn đề xuất chia nhỏ DNN và gán cho CPU/GPU cục bộ để tăng tốc. Tổng quan chi tiết về tăng tốc phần cứng đã có trong khảo sát của Sze et al. [89], do đó bài báo này tập trung vào giải pháp phần mềm.

Bảng 1.1: Tổng hợp các nghiên cứu liên quan đến tăng tốc mô hình học sâu trên thiết bị biên

Nghiên cứu	Mô hình học sâu	Ứng dụng	Thiết bị biên	Phương pháp tối ưu
Taylor [72]	MobileNet [47]	Phân loại hình ảnh	NVIDIA Jetson TX2	lựa chọn mô hình
DeepIoT [73]	LeNet5 [74], VGGNet, BiLSTM [75],	Nhận dạng văn bản, hình ảnh, giọng nói	Intel Edison computing platform	cắt tỉa mô hình (model pruning)
Lai [57]	7-layer CNN	phân loại ảnh	Arm Cortex-M	lượng tử hóa mô hình (model quantization)

ESE [58]	LSTM [76]	nhận dạng giọng nói	XCKU060 FPGA	cắt tỉa và lượng tử hóa mô hình
Bhattacharya [59]	AlexNet [77], VGGNet [78]	nhận dạng giọng nói, hình ảnh	Qualcomm Snapdragon 400/Nvidia Tegra K1/ARM Cortex M0 and M3	làm thưa mô hình (model sparsification)
Adadeep [60]	LeNet, AlexNet, and VGGNet	phân loại hình ảnh, âm thanh, hành vi	điện thoại thông minh, thiết bị đeo, bảng phát triển, thiết bị nhà thông minh	lựa chọn mô hình
DeepMon [61]	Yolo [49], MatConvNet [79]	phát hiện đối tượng	Samsung Galaxy S7	GPU
RSTensorFlow [69]	24-layer CNN, LSTM	phân loại hình ảnh và cử chỉ tay	Nexus 6 and Nexus 5X	GPU
DeepX [70]	AlexNet	nhận dạng giọng nói, hình ảnh	Qualcomm Snapdragon 800/Nvidia Tegra K1	bộ xử lý không đồng nhất (heterogeneous processors)

1.4. Tổng quan về chắt lọc tri thức

1.4.1. Giới thiệu về chắt lọc tri thức

Chắt lọc tri thức là một kỹ thuật tiên tiến trong lĩnh vực học máy, nhằm mục tiêu chuyển giao kiến thức từ một mô hình lớn đã được huấn luyện tốt (được gọi là *mô hình giáo viên*) sang một mô hình nhỏ hơn, hiệu quả hơn (gọi là *mô hình học trò*). Quá trình này đặc biệt quan trọng trong những trường hợp không thể triển khai các mô hình lớn do giới hạn về tài nguyên, chẳng hạn như trên thiết bị di động hoặc hệ thống nhúng. KD cho phép mô hình học trò đạt được hiệu suất gần tương đương với mô hình giáo viên, đồng thời giảm yêu cầu về tính toán và bộ nhớ. Khái niệm này được giới thiệu bởi Hinton et al. vào năm 2015 [80], tận dụng các đầu ra (output) của mô hình giáo viên để hướng dẫn việc huấn luyện mô hình học trò. Phương pháp này đã thu hút nhiều sự quan tâm trong cộng đồng nghiên cứu, đóng góp đáng kể vào các kỹ thuật nén mô hình và suy luận hiệu quả.

Ứng dụng chính của KD không chỉ dừng lại ở việc nén mô hình. KD còn được sử dụng trong các tình huống dữ liệu huấn luyện bị hạn chế hoặc tốn kém để thu thập. Bằng cách tận dụng mô hình giáo viên đã huấn luyện trước đó, mô hình học trò có thể học hiệu quả từ các dự đoán của thầy, từ đó cải thiện khả năng tổng quát hóa. Ngoài ra, KD rất hữu ích trong việc chuyển giao kiến thức chuyên ngành từ mô hình được huấn luyện trên tập dữ liệu lớn, đa dạng sang mô hình học trò được tinh chỉnh cho một tác vụ cụ thể. Tính linh hoạt này khiến KD trở thành một công cụ mạnh mẽ trong nhiều ứng dụng như xử lý ngôn ngữ tự nhiên, thị giác máy tính và hệ thống tự hành, nơi yêu cầu mô hình vừa chính xác vừa hiệu quả.

Trọng tâm của kiến thức chắt lọc là độ lệch Kullback-Leibler (KL

divergence**)** – một thước đo thống kê dùng để định lượng sự khác biệt giữa hai phân phối xác suất. Trong bối cảnh KD, KL divergence đóng vai trò quan trọng trong việc đồng bộ hóa phân phối đầu ra của mô hình giáo viên và mô hình học trò. Về mặt toán học, KL divergence giữa hai phân phối xác suất P (thầy) và Q (trò) được định nghĩa như sau:

$$D_{KL}(P \parallel Q) = \sum_{i} P(i) \log \left(\frac{P(i)}{Q(i)} \right) (1)$$

Công thức trên tính toán độ lệch tương đối (relative entropy), cung cấp một chỉ số về việc một phân phối xác suất lệch bao nhiêu so với phân phối kỳ vọng. Trong thực tế, việc tối thiểu hóa KL divergence giúp đảm bảo rằng phân phối đầu ra của mô hình học trò tiệm cận phân phối của mô hình giáo viên. Sự đồng bộ này thường được đạt được bằng cách kết hợp hàm mất mát entropy chéo (cross-entropy loss) – đo độ chính xác của mô hình học trò – với KL divergence – khuyến khích mô hình học trò mô phỏng hành vi của mô hình giáo viên.

Hàm mục tiêu (objective function) của KD có thể được biểu diễn như sau:

$$\mathcal{L} = \mathcal{L}_{\text{sup}}(y_i, z_i^S) + \lambda \cdot \mathcal{L}_{\text{distill}}(z_i^T, z_i^S) (2)$$

Trong đó, thành phần đầu tiên biểu diễn hàm mất mát có giám sát giữa nhãn thật và dự đoán của mô hình học trò, thành phần thứ hai là mất mát chắt lọc giữa đầu ra của mô hình giáo viên và học trò. Siêu tham số λ (lambda) được dùng để cân bằng mức độ ảnh hưởng của hai thành phần mất mát này. Bằng cách điều chỉnh hợp lý λ, mô hình học trò có thể học tốt cả từ nhãn thực tế và từ mô hình giáo viên, từ đó đạt được hiệu suất và khả năng khái quát hóa tốt hơn.

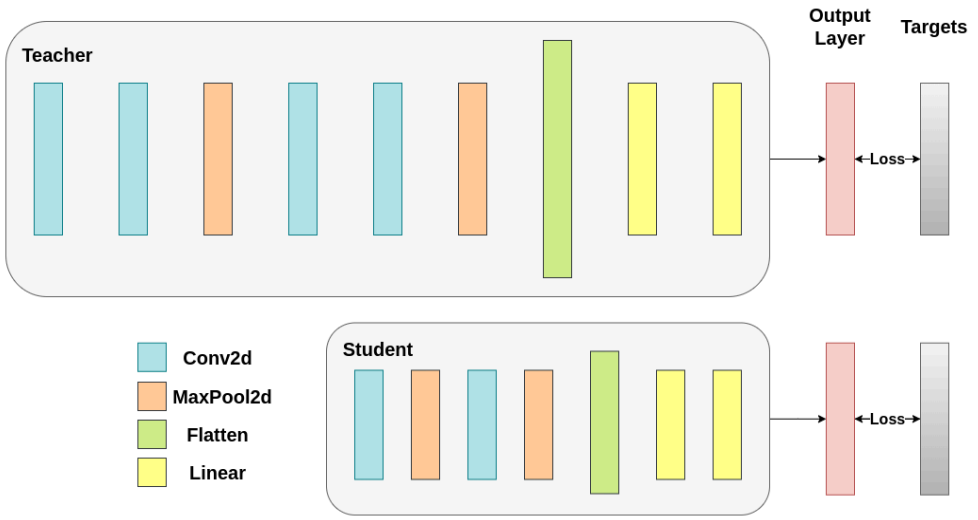
1.4.2. Các nghiên cứu liên quan về chắt lọc tri thức

Tôi khảo sát 13 phương pháp KD tiên tiến, cùng với hai phương pháp mới là Relational Representation Distillation (RRD) và Invariant Consistency Distillation (ICD). Dưới đây là tóm tắt các phương pháp:

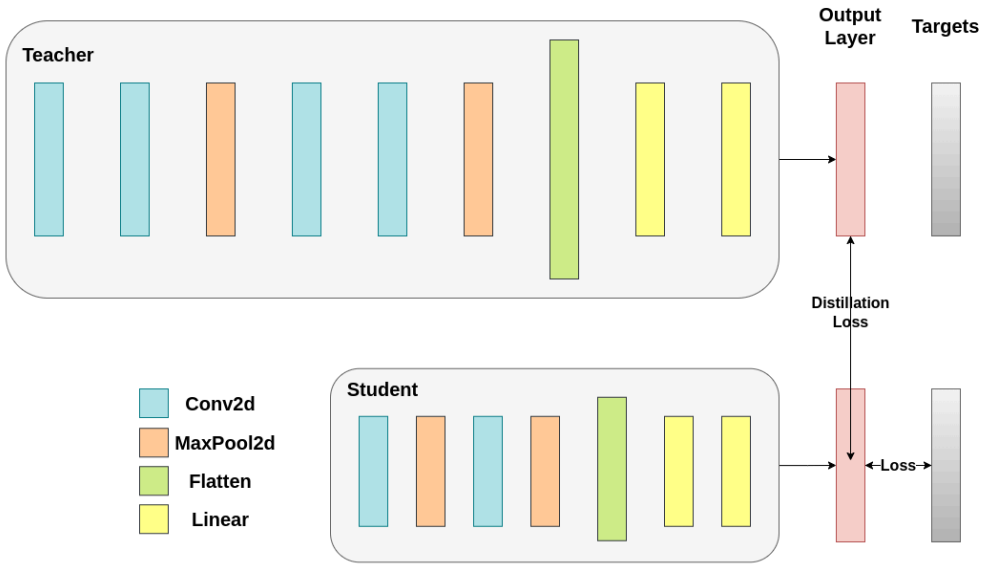
- KD (Hinton et al.) [80]: Phương pháp nền tảng, sử dụng các "nhãn mềm" (soft labels) từ mô hình giáo viên để huấn luyện mô hình học trò bằng cách tối thiểu hóa *Kullback-Leibler divergence* giữa phân phối đầu ra của hai mô hình.
- FitNet [81]: Sử dụng các tầng trung gian của mô hình giáo viên làm tín hiệu hướng dẫn để học trò học các biểu diễn đặc trưng trung gian.
- Paying More Attention to Attention [82]: Truyền bản đồ chú ý (*attention maps*) từ mô hình giáo viên sang học trò, giúp học trò tập trung vào các vùng dữ liệu quan trọng giống như thầy.
- Similarity-Preserving KD [83]: Duy trì tính tương đồng giữa các đặc trưng (feature similarity) của mô hình giáo viên và học trò nhằm bảo toàn thông tin quan hệ giữa các mẫu.
- Correlation Congruence for KD [84]: Căn chỉnh ma trận tương quan giữa các biểu diễn của thầy và trò để truyền đạt cấu trúc tương quan giữa các đặc trưng.
- Variational Information Distillation for Knowledge Transfer [85]: Tối đa hóa thông tin tương hỗ (*mutual information*) giữa biểu diễn của thầy và trò bằng suy luận biến phân.
- Relational KD [86]: Tập trung vào việc truyền tải cấu trúc quan hệ giữa

các điểm dữ liệu (relational knowledge) từ thầy sang trò.

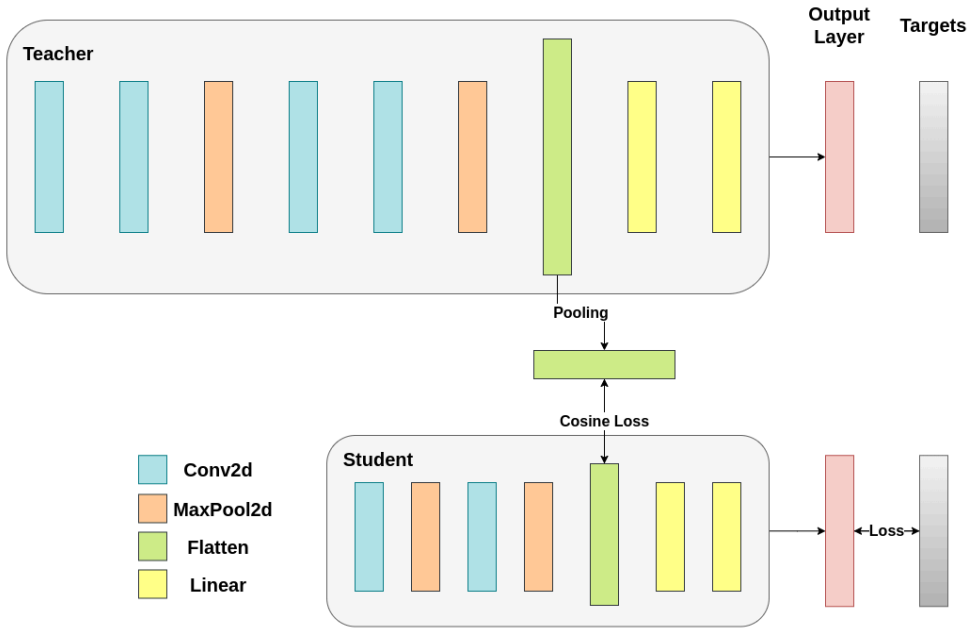
- Probabilistic Knowledge Transfer for Deep Representation Learning [87]: Sử dụng khung xác suất để truyền kiến thức bằng cách khớp phân phối xác suất của đặc trưng giữa thầy và trò.
- Activation Boundaries Formed by Hidden Neurons [88]: Hướng dẫn mô hình học trò bắt chước ranh giới kích hoạt (activation boundaries) được hình thành bởi các neuron ẩn của mô hình giáo viên.
- Factor Transfer [89]: Trích xuất các thành phần nhân tố (factors) từ biểu diễn của mô hình giáo viên và truyền chúng sang mô hình trò, giúp đơn giản hóa quá trình truyền tải.
- Flow of Solution Procedure [90]: Căn chỉnh "quá trình giải quyết" giữa hai mô hình, tức là mô hình trò được huấn luyện để đi theo lộ trình suy luận giống như thầy.
- Neuron Selectivity Transfer [91]: Truyền mô hình kích hoạt neuron (neuron selectivity) từ thầy sang trò nhằm bắt chước cơ chế kích hoạt ra quyết định.
- Contrastive Representation Distillation [92]: Kết hợp học tương phản (*contrastive learning*) trong chắt lọc, giúp học trò học các biểu diễn đặc trưng phân biệt tốt hơn.



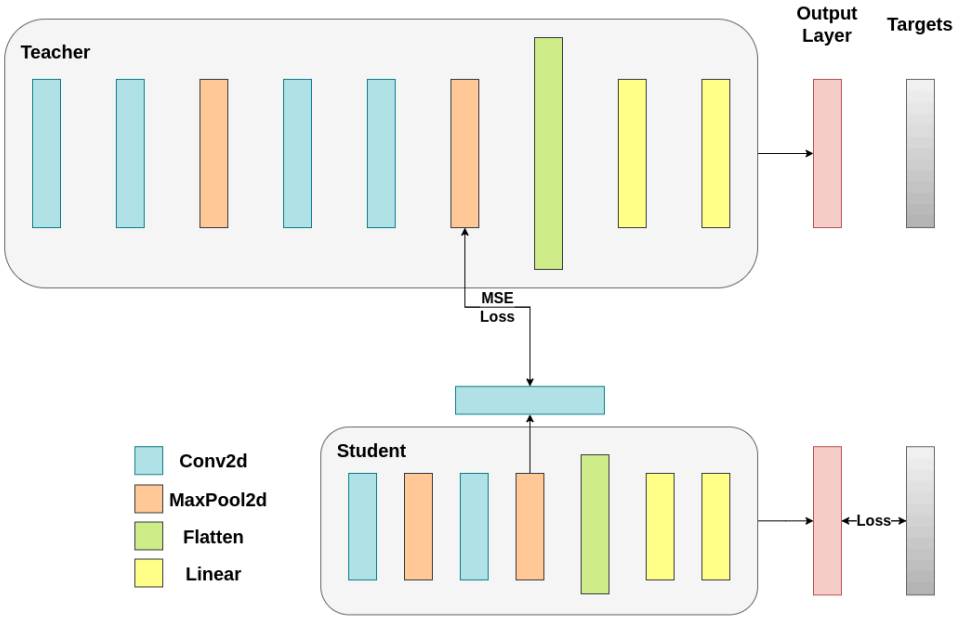
Hình 1.1: Mô hình Student và Teacher được huấn luyện riêng biệt bằng loss giữa output và ground truth. Không có chắt lọc tri thức



Hình 1.2: Mô hình Student học theo output (logits) của Teacher bằng cách tối ưu KDL giữa hai đầu ra



Hình 1.3: Student học theo đặc trưng trung gian (intermediate features) của Teacher sau bước pooling thông qua cosine loss



Hình 1.4: Mô hình Student mô phỏng feature maps của Teacher ở lớp convolution sử dụng MSE loss

So sánh thực nghiệm trên tập CIFAR-100 cho thấy các phương pháp RRD và ICD consistently outperform các phương pháp truyền thống như FitNet, AT, SP trong hầu hết các cặp mô hình giáo viên-trò; khi kết hợp với phương pháp KD gốc (RRD+KD, ICD+KD), hiệu suất còn tăng mạnh hơn, vượt qua cả các phương pháp tiên tiến như CRD+KD; đáng chú ý, trong một số trường hợp, mô hình học trò huấn luyện bằng RRD+KD hoặc ICD+KD còn vượt qua cả mô hình giáo viên, cho thấy khả năng tổng quát hóa vượt trội; so với CRD – một trong những phương pháp mạnh nhất hiện nay – cả RRD và ICD vẫn đạt độ chính xác tương đương hoặc cao hơn, chứng minh tính hiệu quả của các nguyên lý tính nhất quán quan hệ (relational consistency) và tính bất biến (invariance consistency) trong KD.

Teacher Student	wrn-40-2 wrn-16-2	wrn-40-2 wrn-40-1	resnet56 resnet20	resnet110 resnet20	resnet110 resnet32	resnet32x4 resnet8x4	vgg13 vgg8
Teacher Student	75.61 73.26	75.61 71.98	72.34 69.06	74.31 69.06	74.31 71.14	79.42 72.50	74.64 70.36
KD	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet	73.58	72.24	69.21	68.99	71.06	73.50	71.02
AT	74.08	72.77	70.55	70.22	72.31	73.44	71.43
SP	73.83	72.43	69.67	70.04	72.69	72.94	72.68
CC	73.56	72.21	69.63	69.48	71.48	72.97	70.71
VID	74.11	73.30	70.38	70.16	72.61	73.09	71.23
RKD	73.35	72.22	69.61	69.25	71.82	71.90	71.48
PKT	74.54	73.45	70.34	70.25	72.61	73.64	72.88
AB	72.50	72.38	69.47	69.53	70.98	73.17	70.94
FT	73.25	71.59	69.84	70.22	72.37	72.86	70.58
FSP	72.91	N/A	69.95	70.11	71.89	72.62	70.23
NST	73.68	72.24	69.60	69.53	71.96	73.30	71.53
CRD	75.48	74.14	71.16	71.46	73.48	75.51	73.94
CRD+KD	75.64	74.38	71.63	71.56	73.75	75.46	74.29
ICD	74.92	73.69	71.18	71.00	73.11	74.23	72.98
ICD+KD	76.06	74.76	71.81	71.57	73.62	74.99	73.83
RRD	75.01	73.55	70.71	70.72	73.10	74.21	73.99
RRD+KD	75.66	73.77	71.72	71.62	73.48	74.86	74.32

Hình 1.5: So sánh các phương pháp chắt lọc tri thức trên cùng một kiến trúc mô hình

Teacher Student	vgg13 MobileNetV2	ResNet50 MobileNetV2	ResNet50 vgg8	resnet32x4 ShuffleNetV1	resnet32x4 ShuffleNetV2	wrn-40-2 ShuffleNetV1
Teacher Student	74.64 64.60	79.34 64.60	79.34 70.36	79.42 70.50	79.42 71.82	75.61 70.50
KD	67.37	67.35	73.81	74.07	74.45	74.83
FitNet	64.14	63.16	70.69	73.59	73.54	73.73
AT	59.40	58.58	71.84	71.73	72.73	73.32
SP	66.30	68.08	73.34	73.48	74.56	74.52
CC	64.86	65.43	70.25	71.14	71.29	71.38
VID	65.56	67.57	70.30	73.38	73.40	73.61
RKD	64.52	64.43	71.50	72.28	73.21	72.21
PKT	67.13	66.52	73.01	74.10	74.69	73.89
AB	66.06	67.20	70.65	73.55	74.31	73.34
FT	61.78	60.99	70.29	71.75	72.50	72.03
NST	58.16	64.96	71.28	74.12	74.68	74.89
CRD	69.73	69.11	74.30	75.11	75.65	76.05
CRD+KD	69.94	69.54	74.58	75.12	76.05	76.27
ICD	68.22	67.39	73.85	74.07	75.23	74.98
ICD+KD	69.37	69.28	73.88	75.27	76.53	76.39
RRD	67.93	68.84	74.01	74.11	74.64	74.98
RRD+KD	69.98	69.13	74.26	74.78	75.78	76.31

Hình 1.6: So sánh các phương pháp chắt lọc tri thức trên các kiến trúc mô hình khác nhau

1.5. Dữ liệu té ngã

Bảng 1.2 so sánh năm bộ dữ liệu té ngã phổ biến dựa trên tính công khai, đa dạng bối cảnh, số lượng người tham gia, số lớp và số video. Tất cả đều được công khai. GMDCSA24 và FallVision có độ đa dạng cao với nhiều lớp và video. FallVision có số video lớn nhất (6,864), trong khi CAUCAFall có ít bối cảnh nhất. UPFall và URFall cân bằng về số lớp và khung cảnh.

CAUCAFall [93]: Được ghi hình trong môi trường gia đình thực tế với điều kiện ánh sáng và góc nhìn thay đổi; không đa dạng về khung cảnh nhưng có sự đa dạng về người; gồm 10 lớp (5 fall, 5 ADL), tổng 160 video; phù hợp với nghiên cứu mô phỏng điều kiện đời sống thật. | GMDCSA24 [94]: Đa dạng nhất về số lớp (32) và có nhiều khung cảnh, nhiều người tham gia; tuy số video không quá nhiều (162), nhưng rất phù hợp để huấn luyện và đánh giá mô hình đa phân lớp. | FallVision [95]: Có số lượng video lớn nhất (6,864) nhưng chỉ gồm 4 lớp; đa dạng về khung cảnh và người; lý tưởng để huấn luyện mô hình deep learning yêu cầu dữ liệu lớn. | URFall [96]: Gồm 2 lớp với 100 video; có nhiều người và khung cảnh; thường dùng như bộ dữ liệu benchmark cơ bản để đánh giá phương pháp mới. | UPFall [97]: Cân bằng tốt giữa số lớp (11), số video (1,007), người và bối cảnh; phù hợp cho cả mô hình học máy truyền thống và hiện đại; thích hợp để phân biệt nhiều loại hành vi té ngã và hoạt động thường ngày.

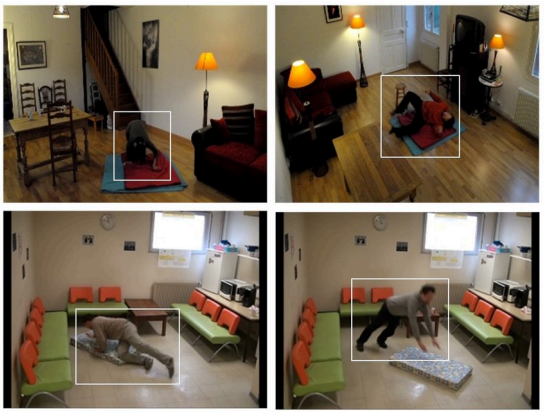
Bảng 1.2: So sánh các bộ data được sử dụng trong bài toán nhận diện té ngã

Bộ dữ liệu	Được đăng công khai (Public dataset)	Nhiều khung cảnh (Cross-background	Nhiều Người (Cross-person)	Số lớp	Số video
CaucaFall [93]	Có	Không	Có	10	160
GMDCSA24 [94]	Có	Có	Có	32	162
FallVision [95]	Có	Có	Có	4	6,864
URFall [96]	Có	Có	Có	2	100
UPFall [97]	Có	Có	Có	11	1,007

Hình 1.7: Minh họa bộ dữ liệu CaucaFall

Hình 1.8: Minh họa bộ dữ liệu GMDCSA24

Hình 1.9: Minh họa bộ dữ liệu FallVision



Hình 1.10: Minh họa bộ dữ liệu URFall

(a) (b) (c)

Hình 1.11: Minh họa bộ dữ liệu UPFall

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT & CÁC NGHIÊN CỨU LIÊN QUAN

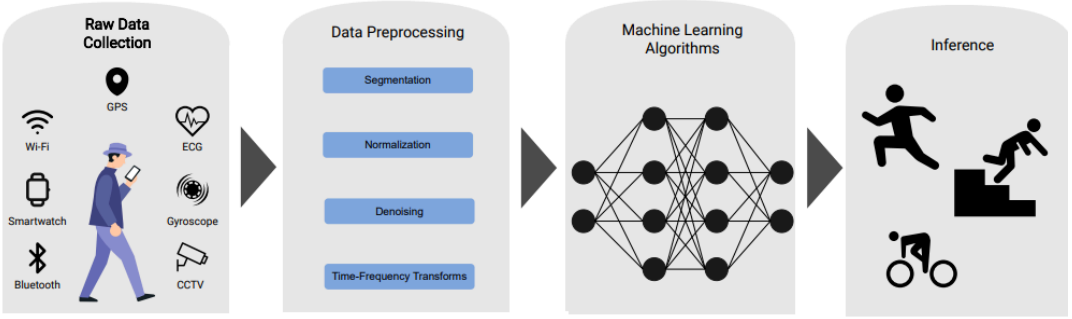
2.1. Bài toán nhận diện hành động (Human Action Recognition - HAR)

Nhận dạng hoạt động của con người (HAR – Human Activity Recognition) đề cập đến quá trình nhận diện và phân loại có hệ thống các hoạt động được thực hiện bởi con người dựa trên dữ liệu thu thập từ nhiều loại cảm biến khác nhau [98]. Đây là một lĩnh vực nghiên cứu liên ngành giao thoa giữa khoa học máy tính, kỹ thuật và khoa học dữ liệu nhằm giải mã các mẫu tín hiệu trong dữ liệu cảm biến và liên hệ chúng với những chuyển động hoặc hành vi cụ thể của con người. Tầm quan trọng của HAR ngày càng được thể hiện rõ nét trong bối cảnh các thiết bị đeo, cảm biến di động và môi trường Internet vạn vật (IoT) phát triển mạnh mẽ. Việc nhận dạng chính xác hoạt động của con người không chỉ nâng cao trải nghiệm người dùng và khả năng tự động hóa, mà còn hỗ trợ trong nhiều ứng dụng đòi hỏi hiểu biết về hành vi con người.

HAR có nhiều ứng dụng đa dạng, bao gồm theo dõi sức khỏe, nhà thông minh, giám sát an ninh, phân tích thể thao và tương tác người–máy. Ví dụ, trong lĩnh vực chăm sóc sức khỏe, HAR có thể hỗ trợ theo dõi từ xa người cao tuổi hoặc bệnh nhân mắc bệnh mãn tính, giúp can thiệp kịp thời và giảm tái nhập viện. Tương tự, trong các ngôi nhà thông minh, việc nhận dạng các hoạt động thường ngày giúp tiết kiệm năng lượng, tăng sự thoải mái và cải thiện an toàn. Trong lĩnh vực thể thao, HAR hỗ trợ vận động viên hoàn thiện kỹ thuật và tư thế, cung cấp phản hồi theo thời gian thực. Ngoài ra, trong giám sát an ninh, các hành vi bất thường có thể được phát hiện kịp thời, đảm bảo phản ứng nhanh chóng. Tóm lại, HAR có tiềm năng mang tính chuyển đổi trên nhiều lĩnh vực, góp phần xây dựng một môi trường phản hồi tốt và trực quan hơn [99].

Một ví dụ về HAR là việc sử dụng các thiết bị đeo như đồng hồ thông minh hoặc thiết bị theo dõi sức khỏe để giám sát và phân loại hoạt động người dùng, như được minh họa trong Hình 2.1 Các thiết bị này thường tích hợp nhiều loại cảm biến, phổ biến nhất là cảm biến gia tốc (accelerometer) và con quay hồi chuyển (gyroscope), lần lượt đo gia tốc tuyến tính và vận tốc góc. Dữ liệu cảm biến thô được thu thập theo các khoảng thời gian định trước, tạo thành một tập dữ liệu chuỗi thời gian [100–103]. Ví dụ, cảm biến gia tốc tạo ra dữ liệu ba trục tương ứng với gia tốc theo các trục x, y và z. Trước khi trích xuất đặc trưng, thường tiến hành một số bước tiền xử lý. Trong các ứng dụng HAR tiêu chuẩn, việc lựa chọn đặc trưng đầu vào và chuẩn hóa dữ liệu là yếu tố then chốt quyết định hiệu suất của thuật toán học máy [104]. Dữ liệu chuỗi thời gian thô thường nhiễu và có thể chứa những dao động không liên quan, vì vậy cần được lọc, thường bằng bộ lọc thông thấp để loại bỏ nhiễu tần số cao [105]. Sau đó, dòng

dữ liệu liên tục được phân đoạn thành các cửa sổ chồng lắp nhau, mỗi cửa sổ đại diện cho một khoảng thời gian nhất định (ví dụ: 2,56 giây với 50% chồng lắp). Kỹ thuật phân đoạn này giúp trích xuất các đặc trưng cục bộ phản ánh rõ ràng mô hình hoạt động. Với mỗi cửa sổ, nhiều đặc trưng được tính toán. Các đặc trưng miền thời gian như giá trị trung bình, phương sai, độ lệch chuẩn và tương quan giữa các trục là phổ biến. Các thuật toán truyền thống như máy vector hỗ trợ (SVM), cây quyết định, rừng ngẫu nhiên và k-láng giềng gần (k-NN) đã được áp dụng rộng rãi. Tuy nhiên, với sự phát triển của học sâu, các phương pháp như mạng nơ-ron tích chập (CNN) và mạng nơ-ron hồi tiếp (RNN) cũng trở nên nổi bật nhờ khả năng mô hình hóa mối quan hệ thời gian phức tạp. Sau khi huấn luyện, mô hình có thể phân loại dữ liệu đầu vào thành các lớp hoạt động đã định sẵn như đi bộ, chạy, ngồi hoặc đứng. Độ chi tiết và chính xác của việc phân loại phụ thuộc vào chất lượng dữ liệu, đặc trưng được trích xuất và hiệu quả của thuật toán được lựa chọn. Tóm lại, HAR sử dụng thiết bị đeo bao gồm một quy trình có hệ thống từ thu thập dữ liệu cảm biến thô đến phân loại hoạt động, dựa trên các phương pháp và thuật toán tính toán tiên tiến.



Hình 2.1. Quy trình tổng quát của các tác vụ nhận dạng hoạt động con người. [109]

Trong một khảo sát được thực hiện bởi [105] về HAR dựa trên thị giác máy tính, các tác giả đã chỉ ra những phát hiện từ các khảo sát liên quan đến chủ đề HAR. Cụ thể, thống kê các khảo sát liên quan từ năm 2010 đến 2019 cho thấy phần lớn các khảo sát tập trung mô tả chi tiết một số khía cạnh cụ thể của HAR [105–109], thay vì cung cấp cái nhìn tổng quan rộng hơn về chủ đề này. Tôi cho rằng để người đọc nắm được xu hướng gần đây trong lĩnh vực này, cần nhiều công trình hơn mô tả phổ rộng các phương pháp và thiết lập được sử dụng trong từng nhánh của HAR. Bài báo này lấp đầy khoảng trống trong thư mục tài liệu bằng cách cung cấp cái nhìn tổng quan về nhiều phương pháp HAR, sử dụng nhiều loại cảm biến và phương thức khác nhau, giúp người đọc xác định được những lỗ hổng còn lại trong nghiên cứu.

Với phạm vi ứng dụng rộng lớn của HAR khi kết hợp các kỹ thuật học máy, tôi đã tổ chức khảo sát theo hướng phân biệt rõ ràng giữa các phương pháp dựa trên cảm biến và các phương pháp dựa trên thị giác. Một đóng góp nổi bật khác của bài viết là việc sử dụng các mô hình ngôn ngữ lớn (LLM) để trích xuất từ khóa liên quan và trả lời câu hỏi, từ đó hỗ trợ việc sắp xếp và lọc cơ sở dữ liệu bài báo một cách hiệu quả.

2.2. Bài toán ước lượng tư thế cơ thể người (Human Pose Estimation - HPE)

Ước lượng tư thế con người (Human Pose Estimation - HPE) là một chủ đề đã được nghiên cứu rộng rãi trong lĩnh vực thị giác máy tính. Nó liên quan đến việc ước lượng cấu hình của các bộ phận cơ thể người từ dữ liệu đầu vào được thu nhận từ các cảm biến, đặc biệt là từ hình ảnh và video. HPE cung cấp thông tin hình học và chuyển động của cơ thể người, và đã được ứng dụng vào nhiều lĩnh vực khác nhau như tương tác người–máy, phân tích chuyển động, thực tế tăng cường (AR), thực tế ảo (VR), y tế, v.v.

Với sự phát triển nhanh chóng của các giải pháp học sâu trong những năm gần đây, các phương pháp này đã chứng minh khả năng vượt trội so với các phương pháp thị giác máy tính truyền thống trong nhiều tác vụ như phân loại ảnh [110], phân đoạn ngữ nghĩa [111] và phát hiện đối tượng [112]. Những tiến bộ đáng kể và hiệu suất ấn tượng đã đạt được khi áp dụng học sâu vào các bài toán HPE. Tuy nhiên, vẫn còn nhiều thách thức cần vượt qua như hiện tượng che khuất (occlusion), thiếu dữ liệu huấn luyện và sự mơ hồ về độ sâu (depth ambiguity).

Đối với HPE 2D từ hình ảnh và video với nhãn tư thế 2D, việc thực hiện là khá dễ dàng và các mô hình học sâu đã đạt hiệu suất rất cao trong ước lượng tư thế của một người đơn lẻ. Gần đây, sự chú ý chuyển hướng sang các bài toán khó hơn như HPE đa người trong những bối cảnh phức tạp với mức độ che khuất cao.

Ngược lại, với HPE 3D, việc thu thập nhãn tư thế 3D chính xác khó khăn hơn nhiều so với nhãn 2D. Các hệ thống ghi lại chuyển động (motion capture systems) có thể thu thập nhãn 3D trong môi trường phòng lab có kiểm soát; tuy nhiên, chúng có những hạn chế khi áp dụng trong môi trường thực tế (in-the-wild). Đối với HPE 3D từ hình ảnh RGB đơn (monocular), thách thức chính nằm ở sự mơ hồ về độ sâu. Trong các thiết lập nhiều góc nhìn (multi-view), việc liên kết giữa các góc nhìn là vấn đề cốt lõi cần được giải quyết.

Một số nghiên cứu đã sử dụng các loại cảm biến như cảm biến độ sâu (depth sensors), đơn vị đo lường quán tính (IMUs), và thiết bị tần số vô tuyến (RF), nhưng các phương pháp này thường không hiệu quả về chi phí và yêu cầu phần cứng chuyên dụng.

Trước tốc độ phát triển nhanh chóng trong nghiên cứu HPE, bài báo này cố gắng theo dõi những tiến bộ gần đây và tổng kết các thành tựu đạt được, nhằm cung cấp một cái nhìn rõ ràng về hiện trạng nghiên cứu trong lĩnh vực ước lượng tư thế con người dựa trên học sâu, cả ở khía cạnh 2D và 3D.

2.3. Chắt lọc tri thức trong HAR

KD đã đóng vai trò then chốt trong việc nén mô hình và suy luận hiệu quả [80]. Phương pháp này cho phép chắt lọc tri thức từ một mô hình giáo viên (teacher) phức tạp, hiệu suất cao sang một mô hình học sinh (student) nhỏ gọn và hiệu quả hơn, trong khi vẫn giữ được mức độ hiệu suất.

Theo thời gian, nhiều cải tiến đã được đề xuất để hoàn thiện KD, tập trung vào các yếu tố như hàm mất mát, rút trích đặc trưng (feature-based distillation), chắt lọc tri thức theo quan hệ (relational knowledge transfer), và học đối kháng (adversarial learning). Khung KD nền tảng do Hinton và cộng sự đề xuất sử dụng các "mục tiêu mềm" (soft targets) được sinh ra từ mô hình giáo viên, giúp mô hình học sinh học được các biểu diễn hiệu quả hơn. Phương pháp này tối thiểu hóa độ lệch Kullback-Leibler (KL divergence) giữa phân phối đầu ra của

giáo viên và học sinh [81, 82, 83].

Bên cạnh việc chắt lọc tri thức trực tiếp, các phương pháp KD có cấu trúc và quan hệ (relational and structural KD) đã được phát triển để nắm bắt các phụ thuộc phức tạp hơn giữa các điểm dữ liệu [87, 88, 89]. Cùng với sự phát triển nhanh chóng của học sâu, các chiến lược sáng tạo đã được đề xuất nhằm nâng cao hiệu quả của KD [87, 88].

Một số nghiên cứu đã áp dụng KD trong nhận dạng hành động. Thoker và Gall [113] đề xuất KD xuyên phương thức (cross-modal distillation) để chuyển tri thức giữa các loại dữ liệu khác nhau. Vu và cộng sự [114] giới thiệu KD tự học (self-knowledge distillation), cho phép mô hình cải thiện chính nó mà không cần mạng giáo viên. Garcia và cộng sự [115] áp dụng KD theo phương thức với các mạng đa dòng (multiple stream networks) để tối ưu hóa hiệu suất. Stroud và cộng sự [117] phát triển D3D – một mạng 3D được rút trích tri thức cho nhận dạng hành động từ video. Liu và cộng sự [117] tập trung vào KD theo từng phần cơ thể (part-level distillation) nhằm nâng cao khả năng nhận dạng hành động từ dữ liệu khung xương (skeleton) chất lượng thấp.

Tuy nhiên, qua nghiên cứu của tôi, tôi nhận thấy rằng dù các phương pháp KD đã được áp dụng rộng rãi trong nhận dạng hành động, nhưng vẫn chưa có nghiên cứu tương tự nào dành riêng cho các bài toán con như phát hiện ngã (fall detection). Do đó, tôi đề xuất áp dụng KD một cách cụ thể cho các tập dữ liệu phát hiện ngã [95, 96, 97, 98, 99].

2.4. Bài toán Nhận diện té ngã (Fall Detection - FD) bằng công nghệ CV

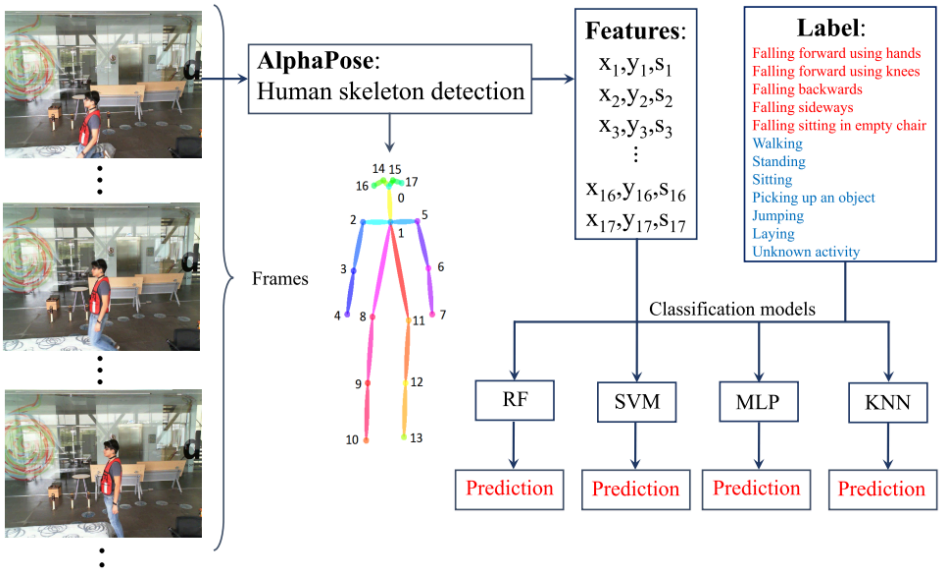
Phát hiện ngã (fall detection) đã được nghiên cứu rộng rãi, chủ yếu thông qua các phương pháp dựa trên thị giác máy tính sử dụng dữ liệu RGB, độ sâu (depth) và hồng ngoại (IR). Trong số đó, dữ liệu RGB là loại được sử dụng phổ biến nhất nhờ tính sẵn có và chi phí thấp [24]. Các phương pháp gần đây tận dụng mô hình học sâu để phân tích đặc trưng không gian–thời gian nhằm phân loại ngã một cách chính xác [25, 26].

Các phương pháp dựa trên khung xương (skeleton-based methods) đã nổi lên như một hướng tiếp cận chủ đạo trong nhận dạng hoạt động con người (HAR) và phát hiện ngã nhờ vào những ưu điểm vốn có, như khả năng biểu diễn đặc trưng hiệu quả và dễ diễn giải [24]. Các phương pháp này thường sử dụng các kỹ thuật ước lượng tư thế cơ thể người (pose estimation) như OpenPose [27] và AlphaPose [28], cho phép trích xuất các đặc trưng không gian và thời gian từ hình ảnh hoặc video.

Trong nhận dạng hoạt động, các mô hình như LSTM và GCN đã được áp dụng để nắm bắt các phụ thuộc tuần tự (sequential dependencies) và quan hệ đồ thị (graph-based dependencies) trong dữ liệu khung xương, từ đó cải thiện hiệu suất trong các bài toán nhận dạng hoạt động động và phức tạp [29, 30].

Nhiều nghiên cứu đã được thực hiện để giảm thời gian xử lý (runtime) của các mô hình phát hiện ngã, từ đó cho phép triển khai thời gian thực trên các thiết bị biên (edge devices). Ví dụ, Ramirez et al. [17] sử dụng các mô hình học máy cơ bản như SVM, KNN, RF và MLP ở lớp phân loại, trong khi Noor et al. [31] áp dụng kỹ thuật lượng hóa hậu huấn luyện (Post-Training Quantization) để giảm độ phức tạp của mô hình, hỗ trợ việc triển khai trên thiết bị biên. Chang et al. [32] tối ưu hóa mô hình bằng cách giảm số lượng tham số để tăng tính tương thích với phần cứng biên.

Tuy nhiên, các hệ thống phát hiện ngã hiện nay vẫn chưa áp dụng những kỹ thuật tối ưu hóa huấn luyện hiện đại nhằm cân bằng giữa việc giảm độ phức tạp mô hình và giữ lại độ chính xác. Xuất phát từ khoảng trống này, tôi đề xuất áp dụng KD để cải thiện các mô hình phát hiện ngã.



Hình 2.2. Quy trình phát hiện té ngã đơn giản dựa trên khung xương người. [17]

2.5. Mô hình YOLO

Con người thông qua vỏ não thị giác – một vùng chính của vỏ não chịu trách nhiệm xử lý thông tin thị giác [118] – có khả năng quan sát, nhận biết [119] và phân biệt giữa các đối tượng gần như tức thì [120]. Việc nghiên cứu cách hoạt động bên trong của vỏ não thị giác và bộ não nói chung đã mở đường cho sự ra đời của các mạng nơ-ron nhân tạo (ANNs) [121] cùng với vô số kiến trúc tính toán thuộc lĩnh vực học sâu.

Trong thập kỷ vừa qua, nhờ những tiến bộ nhanh chóng và mang tính cách mạng trong lĩnh vực học sâu [122], các nhà nghiên cứu đã tập trung vào việc mô phỏng hiệu quả hệ thống thị giác của con người trên máy tính, tức là giúp máy tính có thể phát hiện các đối tượng quan tâm trong ảnh tĩnh và video [123] – một lĩnh vực được gọi là thị giác máy tính (Computer Vision - CV) [124].

CV hiện là một lĩnh vực nghiên cứu phổ biến đối với các nhà nghiên cứu và thực hành học sâu trong thập kỷ này. CV bao gồm các phân ngành như phân loại ảnh [125], phát hiện đối tượng [126], và phân đoạn đối tượng [127]. Cả ba lĩnh vực này đều chia sẻ một kiến trúc chung là sử dụng mạng nơ-ron tích chập (CNNs) [128]. CNNs được coi là tiêu chuẩn mặc định khi làm việc với dữ liệu hình ảnh. So với các phương pháp xử lý ảnh truyền thống và phát hiện thủ công, CNN sử dụng nhiều lớp tích chập kết hợp với các cấu trúc gộp (pooling) nhằm khai thác các đặc trưng ngữ nghĩa sâu ẩn trong từng điểm ảnh của hình ảnh [129].

Trí tuệ nhân tạo (AI) đã tìm thấy cơ hội ứng dụng trong nhiều ngành công nghiệp, từ năng lượng tái tạo [130,131], an ninh, y tế [132] đến giáo dục. Tuy nhiên, một lĩnh vực có tiềm năng lớn để tự động hóa bằng CV là ngành sản xuất. Kiểm tra chất lượng (Quality Inspection - QI) là một phần không thể thiếu trong bất kỳ quy trình sản xuất nào nhằm đảm bảo tính toàn vẹn và sự tin cậy về chất lượng sản phẩm đối với khách hàng [133]. Ngành sản xuất có phạm vi lớn

cho tự động hóa, tuy nhiên, khi kiểm tra bề mặt sản phẩm [134], các khuyết tật có thể có hình thái phức tạp [135], khiến việc kiểm tra bằng con người trở thành một nhiệm vụ khó khăn với nhiều hạn chế như thiên kiến, mệt mỏi, chi phí và thời gian chết [136].

Những hạn chế này tạo điều kiện cho các giải pháp dựa trên CV cung cấp khả năng kiểm tra chất lượng tự động có thể tích hợp vào các quy trình kiểm tra khuyết tật bề mặt hiện tại, từ đó tăng hiệu suất và loại bỏ các nút thắt cổ chai của các phương pháp kiểm tra truyền thống [137].

Tuy nhiên, để thành công, các kiến trúc CV phải tuân theo một bộ yêu cầu triển khai nghiêm ngặt, có thể khác nhau giữa các lĩnh vực sản xuất [138]. Trong phần lớn các ứng dụng, yêu cầu không chỉ là xác định sự tồn tại của khuyết tật mà còn phát hiện nhiều khuyết tật cùng với vị trí cụ thể của từng khuyết tật [139]. Do đó, phát hiện đối tượng (object detection) được ưa chuộng hơn so với phân loại ảnh vì phân loại chỉ đưa ra kết luận có hay không đối tượng mà không cung cấp thông tin vị trí. Các kiến trúc trong lĩnh vực phát hiện đối tượng được chia thành hai loại: bộ phát hiện một giai đoạn (single-stage) và hai giai đoạn (two-stage) [140].

Bộ phát hiện hai giai đoạn chia quá trình phát hiện thành hai bước: trích xuất/đề xuất đặc trưng, sau đó là hồi quy và phân loại để đưa ra đầu ra [141]. Mặc dù điều này có thể mang lại độ chính xác cao, nhưng đi kèm với đó là chi phí tính toán lớn, khiến nó không hiệu quả cho triển khai thời gian thực trên các thiết bị biên có giới hạn. Ngược lại, các bộ phát hiện một giai đoạn kết hợp hai bước này lại thành một, cho phép phân loại và hồi quy trong một lượt duy nhất, từ đó giảm đáng kể yêu cầu tính toán và phù hợp hơn cho các ứng dụng sản xuất thực tế [142].

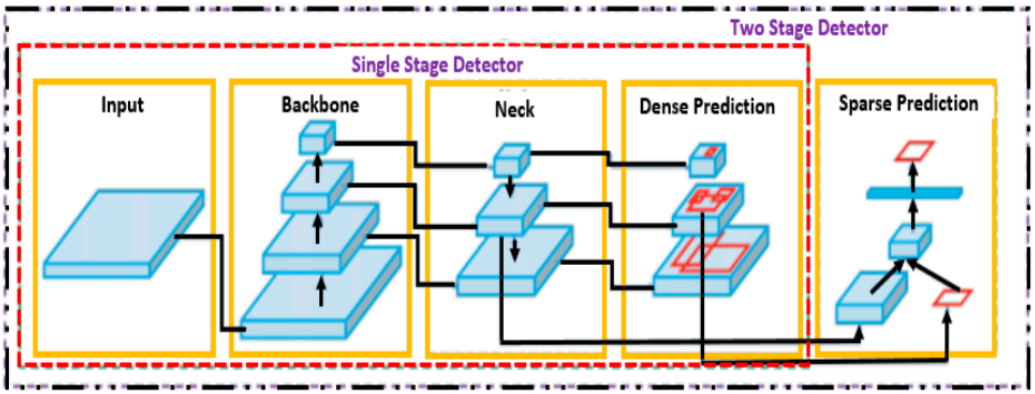
Mặc dù nhiều bộ phát hiện một giai đoạn đã được giới thiệu, chẳng hạn như bộ phát hiện một lần duy nhất (Single Shot Detector - SSD) [143], D-SSD [144], và RetinaNet [145], dòng kiến trúc YOLO (You Only Look Once) [146] đang ngày càng phổ biến nhờ khả năng đáp ứng tốt yêu cầu trong công nghiệp như độ chính xác, nhẹ và dễ triển khai trên thiết bị biên. Nửa sau của thập kỷ qua chứng kiến sự ra đời của nhiều biến thể YOLO, với phiên bản mới nhất là YOLO-v8 ra mắt năm 2022.

CNN có thể được phân loại là mạng nơ-ron truyền thẳng dựa trên tích chập để phục vụ mục đích phân loại [145]. Lớp đầu vào được nối tiếp bởi nhiều lớp tích chập để tạo ra tập các bản đồ đặc trưng ở kích thước nhỏ hơn. Những bản đồ đặc trưng này sau đó được xử lý và chuyển thành vector đặc trưng một chiều, làm đầu vào cho các lớp kết nối đầy đủ. Quá trình trích xuất và xử lý đặc trưng này đóng vai trò then chốt đối với độ chính xác của toàn bộ mạng, do đó việc xếp chồng nhiều lớp tích chập và gộp được thực hiện nhằm thu được các đặc trưng giàu thông tin hơn.

Các kiến trúc phổ biến trong trích xuất đặc trưng bao gồm: AlexNet [146], VGGNet [147], GoogleNet [148], và ResNet [149]. AlexNet được giới thiệu năm 2012, bao gồm năm lớp tích chập, ba lớp gộp, và ba lớp kết nối đầy đủ, chủ yếu dùng cho phân loại ảnh. VGGNet tập trung vào nâng cao hiệu năng bằng cách tăng chiều sâu bên trong mạng, giới thiệu các biến thể như VGG-16/19. GoogleNet giới thiệu khái niệm xếp chồng các module ‘inception’, trong khi ResNet đưa ra khái niệm kết nối tắt (skip-connections) để giữ lại thông tin từ các lớp trước đó và truyền sang các lớp sau.

Mục tiêu của một bộ phát hiện đối tượng là suy ra liệu đối tượng quan tâm có tồn tại trong ảnh hay trong khung hình video hay không. Nếu có, bộ phát hiện sẽ trả về lớp tương ứng và vị trí, tức là kích thước tọa độ của đối tượng. Phát hiện đối tượng được chia thành hai loại: phương pháp hai giai đoạn và một giai đoạn như minh họa trong Hình 2.3. Phương pháp hai giai đoạn bắt đầu bằng việc đề xuất nhiều vùng khả thi, sau đó tiến hành dự đoán trên các vùng này trong giai đoạn thứ hai. Ví dụ về các bộ phát hiện hai giai đoạn bao gồm các biến thể nổi tiếng của R-CNN [150], như Fast R-CNN [151] và Faster R-CNN [152], có độ chính xác cao nhưng hiệu suất tính toán thấp.

Ngược lại, phương pháp một giai đoạn biến nhiệm vụ thành một bài toán hồi quy, loại bỏ giai đoạn chọn vùng ứng viên. Do đó, quá trình chọn và dự đoán vùng ứng viên được thực hiện trong một lượt duy nhất. Các kiến trúc thuộc nhóm này ít đòi hỏi tài nguyên tính toán hơn, tạo ra tốc độ khung hình trên giây (FPS) cao hơn, nhưng thường có độ chính xác kém hơn so với các bộ phát hiện hai giai đoạn.



Hình 2.3. Mô hình đơn giản của bài toán phát hiện vật thể [153]

2.6. Mô hình ST-GCN

Dữ liệu khung xương đã được sử dụng rộng rãi trong nhận diện hành động con người trong những năm gần đây nhờ khả năng biểu diễn cấu trúc tư thế không phụ thuộc vào góc nhìn. So với dữ liệu video RGB, dữ liệu khung xương con người dưới dạng đồ thị là một biểu diễn nhỏ gọn và bền vững hơn cho chuyển động của con người. Việc biểu diễn dữ liệu khung xương dưới dạng đồ thị giúp nó ít nhạy cảm hơn với thay đổi góc nhìn, hiện tượng che khuất, nhiễu hậu cảnh, sự đa dạng về tư thế giữa các lớp, điều kiện ánh sáng và trang phục. Dữ liệu khung xương có thể được thu thập thông qua các thiết bị camera (như Kinect và cảm biến chuyển động) hoặc các thuật toán ước lượng tư thế con người [154]–[156].

Các công trình ban đầu trong lĩnh vực này được truyền cảm hứng từ các kỹ thuật xử lý ảnh, nhằm mô hình hóa hình dạng và sự phụ thuộc về chuyển động của các khớp khung xương bằng cách sử dụng các đặc trưng thủ công. Một số nghiên cứu trước đây đề xuất sử dụng các đặc trưng như Histogram of Oriented Optical Flow (HOF) [157], Histogram of Oriented Gradient (HOG) [158], Speeded Up Robust Feature (SURF) [159] và Scale-Invariant Feature Transform (SIFT) [160] để trích xuất các mối quan hệ có tính phân biệt. Tuy nhiên, các đặc trưng này không có khả năng trích xuất các phụ thuộc không gian-thời gian để bao quát đặc trưng chuyển động và quỹ đạo. Phương pháp

Improved Dense Trajectory (IDT) [161] có khả năng tích hợp các quỹ đạo chuyển động một cách hiệu quả, nhưng vẫn chưa đủ khả năng mô hình hóa các phụ thuộc thời gian mạnh mẽ.

Dù có thể mô hình hóa các phụ thuộc có tính phân biệt, các đặc trưng thủ công lại nhạy cảm với siêu tham số và yêu cầu cách tiếp cận tinh tế trong quá trình mô hình hóa. Các nghiên cứu gần đây sử dụng các kỹ thuật học sâu để tự động hóa việc tạo đặc trưng và trích xuất đặc trưng không gian-thời gian từ chuỗi video. Mạng Neural hồi tiếp (Recurrent Neural Networks - RNN) đạt hiệu suất tốt hơn trong mô hình hóa đặc trưng thời gian [162]–[164], nhưng vốn dĩ không hiệu quả trong việc mô hình hóa các phụ thuộc dài hạn. Mạng Long-Short Term Memory (LSTM) kết hợp với thông tin không gian-thời gian đã mô hình hóa hiệu quả bài toán nhận diện hành động nhờ khả năng xử lý các phụ thuộc dài hạn [165]–[167].

Các nghiên cứu [168], [169] sử dụng mạng Convolutional Neural Network (CNN) cơ bản và mạng tích chập 3 chiều (3D-CNN hay C3D) cho phân loại hành động. Kỹ thuật học tăng cường sâu (Deep Reinforcement Learning) kết hợp với mạng chắt lọc khung hình chính (keyframe distillation network) cũng đã cho thấy hiệu suất tốt hơn trên các bộ dữ liệu chuẩn [170].

Phần lớn các nghiên cứu triển vọng trong lĩnh vực này tập trung vào các mạng GCN dựa trên khung xương. Các phương pháp GCN được đề xuất trong [171]–[173] sử dụng đồ thị dữ liệu dựa trên khung xương và các phép toán tích chập đồ thị để mô hình hóa đặc trưng không gian-thời gian. Sijie Yan và cộng sự [18] là người đầu tiên đề xuất việc sử dụng khung xương con người để xây dựng đồ thị và áp dụng tích chập đồ thị để học đặc trưng cho nhận diện hành động con người.

Mạng tích chập đồ thị (GCN) áp dụng các phép tích chập lên dữ liệu dạng đồ thị, thay vì dữ liệu ảnh như trong các mạng CNN cổ điển [174]. Nhiều nhà nghiên cứu đã sử dụng GCN trong hàng loạt ứng dụng nhờ vào kết quả nổi bật của chúng trên dữ liệu đồ thị. Có hai loại GCN: dạng phổ (spectral) và dạng không gian (spatial). GCN phổ chuyển đổi đồ thị sang miền phổ và áp dụng biến đổi Fourier đồ thị, trong khi GCN không gian trích xuất thông tin từ các nút lân cận. Phương pháp được đề xuất trong nghiên cứu này sử dụng GCN phổ.

Nhận diện hành động con người dựa trên khung xương đã được nghiên cứu rộng rãi bằng cách sử dụng các mạng GCN. Phần lớn các công trình nghiên cứu nổi bật gần đây tập trung vào mạng ST-GCN để khai thác chuyển động và sự phụ thuộc theo thời gian trong một chuỗi video.

Mô hình ST-GCN truyền thống bao gồm một tập hợp các khối ST-GCN áp dụng tích chập đồ thị không gian và thời gian một cách xen kẽ trên đồ thị khung xương [175]. Cuối cùng, các lớp kết nối đầy đủ (fully connected) kết hợp với bộ phân loại SoftMax được sử dụng để dự đoán nhãn hành động.

Zheng Wanqiang và Punan Jing đề xuất mạng tích chập đồ thị không gian-thời gian dựa trên khung xương (ST-GCN) [176]. Dữ liệu khung xương được trích xuất từ video bằng thuật toán ước lượng tư thế con người OpenPose và chuyển thành đồ thị khung xương. ST-GCN sử dụng đồ thị tĩnh cố định với sơ đồ phân vùng ba thành phần để trích xuất đặc trưng không gian và thời gian phục vụ phân loại hành động.

Yang và cộng sự sử dụng phương pháp học dựa trên dữ liệu để tạo đồ thị

khung xương động cùng với ba sơ đồ phân vùng. Các đặc trưng không gian được khai thác thông qua ma trận kề dựa trên cơ chế attention cho mỗi khung hình khung xương. Trong khi đó, các đặc trưng thời gian được trích xuất từ thông tin ngữ nghĩa về vận tốc. Mạng được triển khai dưới dạng mạng GCN tổng quát dựa trên attention (AG-GCN) [177].

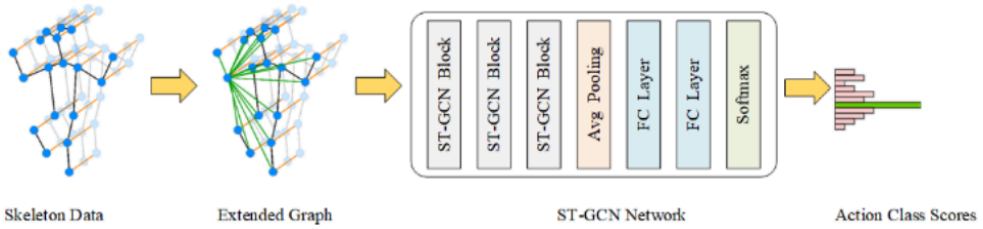
Shi Lei và cộng sự đề xuất một mạng mới mang tên mạng GCN hai dòng thích ứng dựa trên khung xương (2S-AGCN) cho bài toán nhận diện hành động con người từ video [178]. Mạng xử lý đồng thời thông tin bậc nhất và bậc hai. Thông tin bậc nhất đại diện cho vị trí khớp, còn thông tin bậc hai đại diện cho chiều dài và hướng của các xương trong bộ khung con người. Phương pháp này đạt được sự cải thiện đáng kể về hiệu suất trên các tập dữ liệu chuẩn.

Sijie Yan và cộng sự đề xuất một mạng ST-GCN mới cho bài toán phân loại hành động con người bằng cách học đặc trưng không gian và thời gian từ dữ liệu [179]. Mô hình này đạt được cải thiện đáng kể trên hai tập dữ liệu quy mô lớn.

Cheng K và cộng sự đề xuất mạng GCN dịch chuyển (Shift-GCN) mới cho phân loại hành động con người với kiến trúc tiết kiệm tài nguyên tính toán. Các phép toán dịch chuyển đồ thị nhẹ cùng với các phép tích chập điểm giúp toàn bộ mạng trở nên nhẹ và sử dụng ít tài nguyên tính toán. Ngoài ra, phép dịch chuyển đồ thị giúp các trường tiếp nhận thích ứng trong quá trình tích chập không gian và thời gian.

Huang Z và cộng sự đề xuất mạng GCN không gian-thời gian kiểu Inception (Inception ST-GCN). Phương pháp này cải thiện hiệu suất bằng cách trích xuất và tổng hợp thông tin về tỉ lệ và phép biến đổi từ các đường dẫn và cấp độ khác nhau.

Tất cả các phương pháp được đề cập ở trên đều sử dụng đồ thị khung xương dựa trên cấu trúc xương tự nhiên của con người. Ngoài ra, một sơ đồ phân vùng ba phần được sử dụng để chuyển đồ thị có độ dài biến đổi thành tensor có kích thước cố định.

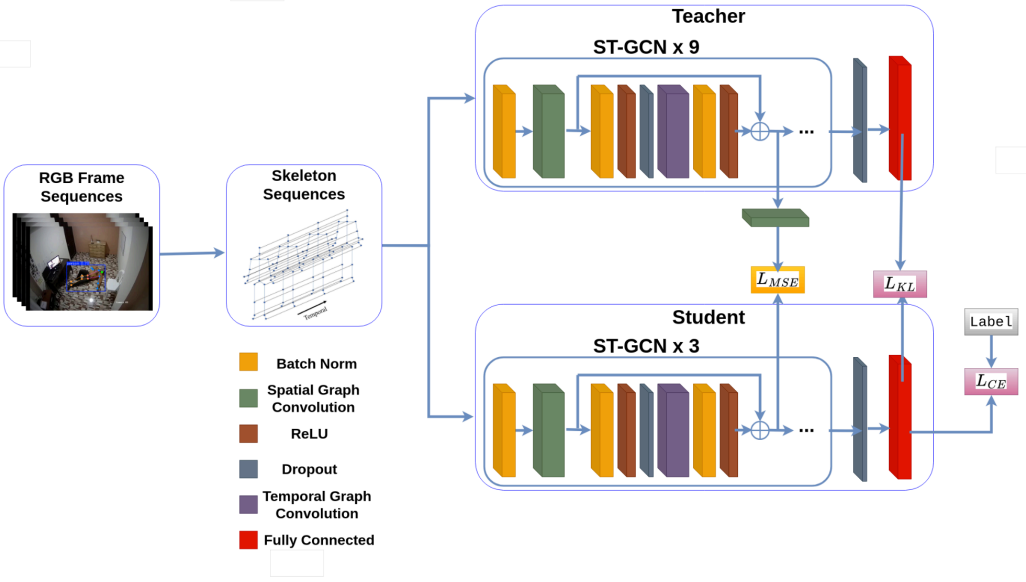


Hình 2.4. Tổng quan hệ thống cho bài toán nhận dạng hành động của con người dựa trên bộ xương sử dụng ST-GCN

CHƯƠNG 3: GIẢI PHÁP ĐỀ XUẤT

Nghiên cứu này tập trung vào việc cải thiện phát hiện té ngã và nhận diện hoạt động bằng cách tận dụng KD cho nhận diện hành động dựa trên khung xương. Giả thuyết chính được đưa ra là: mô hình học sinh có thể đạt hiệu suất tương đương với mô hình giáo viên, trong khi giảm đáng kể độ phức tạp tính toán, từ đó giúp các ứng dụng thời gian thực khả thi hơn. Quy trình làm việc được phát triển cho nghiên cứu này được minh họa trong Hình 3.1. Trong đó, giải pháp cốt lõi là kết hợp các hàm mất mát (loss functions) được đề xuất riêng lẻ trong các nghiên cứu trước [80, 81] thành một chiến lược huấn luyện thống nhất. Mục tiêu của việc kết hợp này là tận dụng điểm mạnh của từng hàm loss để truyền đạt thông tin tri thức từ mô hình giáo viên sang mô hình học sinh một cách hiệu quả hơn, đồng thời vẫn đảm bảo mô hình học sinh duy trì hiệu suất cao với chi phí tính toán thấp.

4.1. Tổng quan kiến trúc hệ thống



Hình 3.1. Tổng quan quy trình mô hình nhận diện té ngã để xuất

Với một chuỗi khung xương được biểu diễn dưới dạng đồ thị $G = (V, E)$, trong đó V biểu thị các khớp cơ thể và E biểu thị các kết nối của bộ xương, ST-GCN được sử dụng để trích xuất các đặc trưng không gian - thời gian từ chuỗi này. Phép toán cốt lõi của ST-GCN được định nghĩa như sau:

$$H^{(l+1)} = \sigma\left(\sum_k \Lambda_k^{-\frac{1}{2}} A_k \Lambda_k^{-\frac{1}{2}} H^{(l)} W_k\right) \quad (3)$$

Trong đó, A_k là ma trận kề đã được chuẩn hóa, mã hóa mối liên kết giữa các khớp xương; $H^{(l)}$ llà ma trận đặc trưng tại tầng l; $W_k$là ma trận trọng số học được;và σlà hàm kích hoạt phi tuyến.

Để chuyển tri thức từ mô hình ST-GCN giáo viên sang mô hình ST-GCN học sinh, phương pháp KD được áp dụng. Quá trình này kết hợp giữa hàm mất mát phân loại và hàm mất mát truyền tri thức.

$$L_{KL} = T^2 \sum_i p_T(y_i) \log \frac{p_T(y_i)}{p_S(y_i)} (4)$$

Trong đó, $p_T(y_i)$và $p_S(y_i)$ lần lượt là xác suất đầu ra của mô hình giáo viên và học sinh, đã được làm mềm bởi một tham số nhiệt độ T:

$$p(y_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} (5)$$

Trong đó z_i là đầu ra logit cho lớp i. Hàm mất mát cross-entropy tiêu chuẩn với nhãn thực được ký hiệu là L_{CE}

Để khuyến khích mô hình học sinh không chỉ sao chép các dự đoán cuối cùng mà còn căn chỉnh các biểu diễn bên trong của nó với mô hình giáo viên, tôi giới thiệu một hàm mất mát đặc trưng ở mức khớp (joint-level features loss) dựa trên sai số bình phương trung bình (Mean Squared Error - MSE). Gọi $f_T(x)$ và $f_S(x)$ lần lượt là các đặc trưng ở mức khớp được trích xuất bởi mạng giáo viên và mạng học sinh tại một lớp trung gian được chọn. Hàm mất mát đặc trưng được định nghĩa như sau:

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \| f_T(x_i) - f_S(x_i) \|_2^2 (6)$$

Trong đóN là kích thước batch, và chuẩn L2 bình phương tương ứng với sai số bình phương trung bình (MSE) giữa các đặc trưng của mô hình giáo viên và mô hình học sinh.

Ba hàm mất mát — cross-entropy (mất mát phân loại), distillation (mất mát chắt lọc tri thức) và feature (mất mát đặc trưng) — được kết hợp thành một hàm mục tiêu duy nhất:

$$L = \alpha L_{KL} + \beta L_{\text{MSE}} + (1 - \alpha - \beta) L_{CE} (7)$$

trong đó α cân bằng giữa mất mát cross-entropy có giám sát và mất mát chắt lọc tri thức KD loss, còn β kiểm soát mức độ bắt chước đặc trưng ở cấp độ bên trong.

Bằng cách tối thiểu hóa hàm mục tiêu này, mô hình học sinh ST-GCN không chỉ học cách tái tạo phân phối đầu ra đã làm mềm của mô hình giáo viên mà còn căn chỉnh các biểu diễn bên trong của mình, từ đó dẫn đến quá trình chắt lọc tri thức hiệu quả và ổn định hơn, đồng thời vẫn duy trì được một mô hình gọn nhẹ.

4.2. Thu thập và xử lý dữ liệu

Để trích xuất đặc trưng khung xương một cách hiệu quả trên các thiết bị

biên, tôi sử dụng YOLOv11 [153] Pose, một mô hình phát hiện điểm khớp nhẹ nhưng chính xác, được tối ưu hóa cho xử lý thời gian thực. Với một khung hình video đầu vào, YOLOv11 [153] Pose phát hiện các khớp của con người và xuất ra một tập hợp các điểm đặc trưng 2D:

$$K = \{(x_i, y_i, c_i) \mid i \in \{1, \ldots, N\}\} (8)$$

trong đó (x_i, y_i)biểu thị tọa độ pixel của điểm khớp thứ i, c_i là điểm tin cậy (confidence score), và Nlà tổng số điểm khớp được phát hiện. Để đảm bảo tính ổn định và nhất quán trên nhiều độ phân giải và góc nhìn camera khác nhau, tôi áp dụng hai kỹ thuật chuẩn hóa: Chuẩn hóa Min-Max và Chuẩn hóa theo chiều dài cơ thể.

Chuẩn hóa Min-Max đưa các điểm khớp vào một khoảng giá trị từ 0 đến 1 nhằm loại bỏ sự phụ thuộc vào độ phân giải tuyệt đối của hình ảnh. Với tọa độ điểm khớp nhỏ nhất và lớn nhất lần lượt là $x_{min}, x_{max}, y_{min}, y_{max}$, mỗi điểm khớp được chuẩn hóa như sau:

$$\hat{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \quad \hat{y}_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} (9)$$

Phép biến đổi này đảm bảo rằng tất cả các điểm khớp đều được biểu diễn theo một tỷ lệ cố định, giúp mô hình tổng quát hóa tốt hơn trên các đối tượng và thiết lập máy ảnh khác nhau. Vì tỷ lệ cơ thể người nhìn chung là tương đối đồng nhất giữa các cá nhân, tôi tiếp tục chuẩn hóa các điểm khớp dựa trên chiều dài cơ thể.

Tôi định nghĩa chiều dài cơ thể L là khoảng cách Euclid giữa cổ K_{neck} và trung điểm của hai khớp hông K_{hip}:

$$L = \|K_{\text{neck}} - K_{\text{hip}}\|_2 (10)$$

Sau đó, mỗi điểm khớp sẽ được chuẩn hóa tương đối theo chiều dài cơ thể này như sau:

$$\tilde{x}_i = \frac{x_i - x_{\text{hip}}}{L}, \quad \tilde{y}_i = \frac{y_i - y_{\text{hip}}}{L} (11)$$

Phương pháp này giúp bù đắp cho sự khác biệt về khoảng cách đến camera và cho phép các đặc trưng trích xuất trở nên ít phụ thuộc vào sự thay đổi về tỉ lệ. Bằng cách tích hợp các kỹ thuật chuẩn hóa này, các điểm khớp được trích xuất trở thành những đặc trưng mạnh mẽ, không phụ thuộc vào độ phân giải, phù hợp cho việc nhận dạng hành động dựa trên khung xương trên các thiết bị biên có tài nguyên hạn chế.

4.3. Đánh giá

Để đánh giá hiệu suất của mô hình, tôi sử dụng phương pháp cross-validation, điều chỉnh chiến lược phân chia dựa trên đặc điểm của bộ dữ liệu. Cụ thể, tôi áp dụng phương pháp cross-validation Leave-One-Subject-Out (LOSO) nhằm đảm bảo khả năng tổng quát hóa đối với các cá nhân chưa từng thấy trong quá trình huấn luyện. Thay vì chia ngẫu nhiên, tôi lặp lại quá trình

huấn luyện và kiểm tra bằng cách sử dụng dữ liệu của một người tham gia làm tập kiểm tra, trong khi dữ liệu của tất cả những người còn lại được dùng để huấn luyện. Gọi S là tập hợp tất cả các đối tượng (subject), và S_i là dữ liệu tương ứng với đối tượng i. Với mỗi lần lặp, tập huấn luyện và tập kiểm tra được xác định như sau:

$$S_{\text{train}} = S \setminus S_i, \quad S_{\text{test}} = S_i (12)$$

Điều này đảm bảo rằng mô hình được đánh giá dựa trên khả năng tổng quát hóa đối với những cá nhân hoàn toàn chưa từng xuất hiện trong quá trình huấn luyện. Chỉ số hiệu suất cuối cùng được tính bằng trung bình của tất cả các vòng lặp.

Để định lượng hiệu suất, tôi sử dụng chỉ số F1-score, là trung bình điều hòa giữa độ chính xác (precision) và độ bao phủ (recall), được định nghĩa như sau:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} (13)$$

Với precision và recall là:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} (14)$$

Tại đây,TP (số dương tính thật), FP(số dương tính giả) và FN(số âm tính giả) được tính dựa trên dự đoán của mô hình. Chỉ số F1-score cung cấp một thước đo độ chính xác cân bằng, đặc biệt hữu ích đối với các tập dữ liệu mất cân bằng, nơi độ chính xác (precision) và độ bao phủ (recall) có thể khác biệt đáng kể.

CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

5.1. Bộ dữ liệu

Tôi đánh giá phương pháp của mình bằng cách sử dụng nhiều bộ dữ liệu công khai khác nhau, bao gồm CaucaFall, GMDCSA-24, FallVision, URFall và UPFall, vốn khác nhau về kích thước mẫu, môi trường và sự đa dạng của đối tượng. Để đảm bảo tính nhất quán và so sánh công bằng, tôi chuẩn hóa tất cả các bộ dữ liệu bằng cách gộp các hành động liên quan đến ngã vào một lớp duy nhất là Fall, và tất cả các hoạt động còn lại vào lớp Non-Fall. Việc phân loại nhị phân này cho phép đánh giá hiệu suất một cách thống nhất trên các bộ dữ liệu khác nhau, bất chấp sự khác biệt trong cách chú thích.

Bảng 4.1: Dữ liệu được dùng trong nghiên cứu

Bộ dữ liệu	Số mẫu	Số đối tượng	Cross-Validation
CaucaFall [93]	100	10	Cross-Person
GMDCSA-24 [94]	162	4	Cross-Person
FallVision [95]	6,864	4	Cross-Person
URFall [96]	100	30	Cross-Background
UPFall [97]	1,007	17	Cross-Person

5.2. Chi tiết triển khai

Trong các thí nghiệm, tôi sử dụng một mô hình sinh viên (student) gồm 3 lớp ST-GCN và hai mô hình giáo viên (teacher): một ST-GCN 4 lớp (với số kênh lần lượt là 64, 128, 256, 512) và một ST-GCN 9 lớp (với các kênh là 64×4, 128×3, 256×3). Mô hình sinh viên gồm 3 khối STGC (số kênh là 8, 16, 32). Tất cả các mô hình được huấn luyện trong 60 epoch sử dụng bộ tối ưu Adam (với learning rate = 0.001, weight decay = 10^{-5}), và bộ điều chỉnh learning rate theo nhiều bước (giảm ở các mốc 30, 40 và 50 epoch với hệ số decay = 0.1).Quá trình huấn luyện sử dụng hàm mất mát cross-entropy có trọng số, và hiệu suất mô hình được đánh giá bằng F1-score. Đối với distillation, tôi sử dụng nhiệt độ T=4, với trọng số cho hàm mất mát soft target là 0.7 và trọng số cho cross-entropy là 0.3. Mô hình sinh viên học từ cả hai mô hình giáo viên thông qua KD. Việc huấn luyện và đánh giá mô hình được thực hiện trên hệ

thống có hỗ trợ GPU.

Bảng 4.2: Danh sách các mô hình sử dụng trong nghiên cứu

Mô hình	Chi tiết các lớp	Số tham số	Thời gian chạy (ms)
Sinh viên (3 lớp)	(8, 16, 32)	17,203	4.003
Giáo viên (4 lớp)	(64, 128, 256, 512)	3,662,286	9.655
Giáo viên (9 lớp)	(64×4, 128×3, 256×3)	3,031,584	7.885

5.3. Kết quả triển khai

Bảng 4.3 cho thấy mô hình sinh viên sau khi áp dụng KD đều cải thiện rõ rệt về hiệu suất. Cụ thể, điểm F1 trung bình tăng đáng kể, với mức tăng điển hình như trên URFall (+0.0703) và CaucaFall (+0.0656). Trong khi đó, giá trị loss có xu hướng giảm trên hầu hết các tập dữ liệu, đáng chú ý là FallVision (–0.1530) và GMDCSA-24 (–0.0827). Điều này khẳng định tính hiệu quả của KD trong việc nâng cao hiệu suất mô hình nhẹ mà không cần tăng số lượng tham số.

Bên cạnh các chỉ số chính xác, tôi cũng so sánh hiệu suất suy luận (inference) của mô hình sinh viên với các mô hình cơ sở khác. Bảng 4.4 hiển thị thời gian suy luận trung bình trên mỗi mẫu và số khung hình trên giây (FPS) tương ứng đạt được trên cùng một phần cứng. Như được trình bày, mô hình sinh viên nhẹ của tôi đạt tốc độ suy luận nhanh hơn đáng kể, khiến nó phù hợp hơn cho các ứng dụng thời gian thực, đặc biệt là trong các môi trường nhúng hoặc hạn chế tài nguyên.

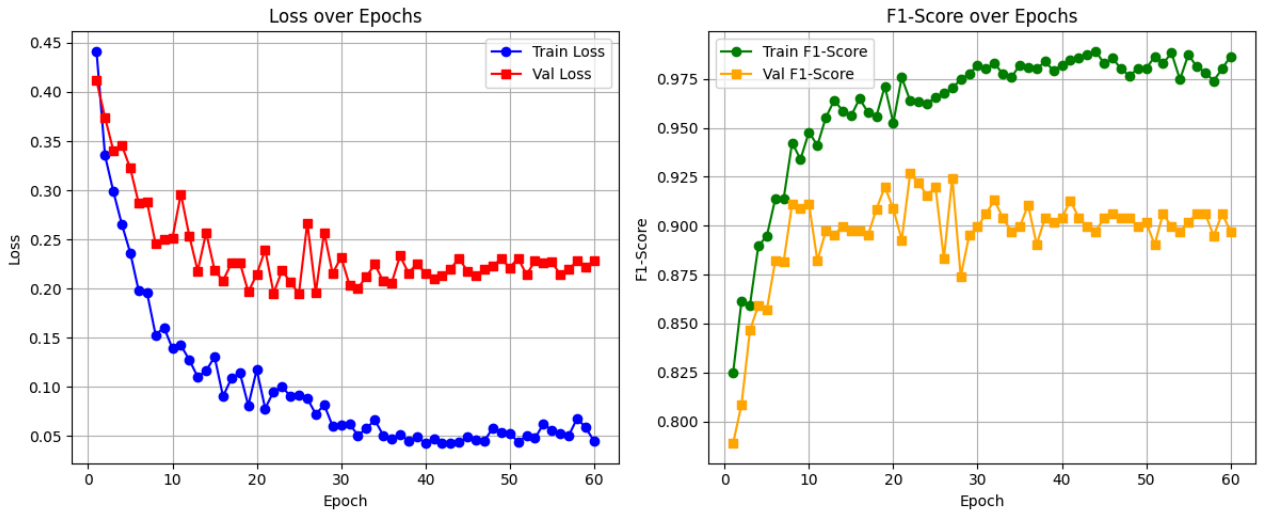
Bảng 4.3. So sánh hiệu suất giữa mô hình sinh viên và mô hình giáo viên trên các tập dữ liệu khác nhau

Bộ dữ liệu	Mô hình sinh viên	Mô hình giáo viên	Mô hình giáo viên		Mô hình sinh viên		Mô hình sinh viên được áp dụng KD	
			Loss	F1	Loss	F1	Loss	F1
CaucaFall [93]	3-layer (8, 16, 32)	4-layer (64, 128, 256, 512)	0.1267	0.9687	0.3553	0.8398	0.3674	0.9054
GMDCSA-24 [94]	3-layer (8, 16, 32)	9-layer (64×4, 128×3, 256×3)	0.2254	0.9302	0.3894	0.8600	0.3067	0.9069
FallVision [95]	3-layer (8, 16, 32)	4-layer (64, 128, 256, 512)	0.3545	0.8802	0.5604	0.8008	0.4074	0.8339
URFall [96]	3-layer (8, 16, 32)	4-layer (64, 128, 256,	0.1323	1.0000	0.3018	0.7868	0.3115	0.8571

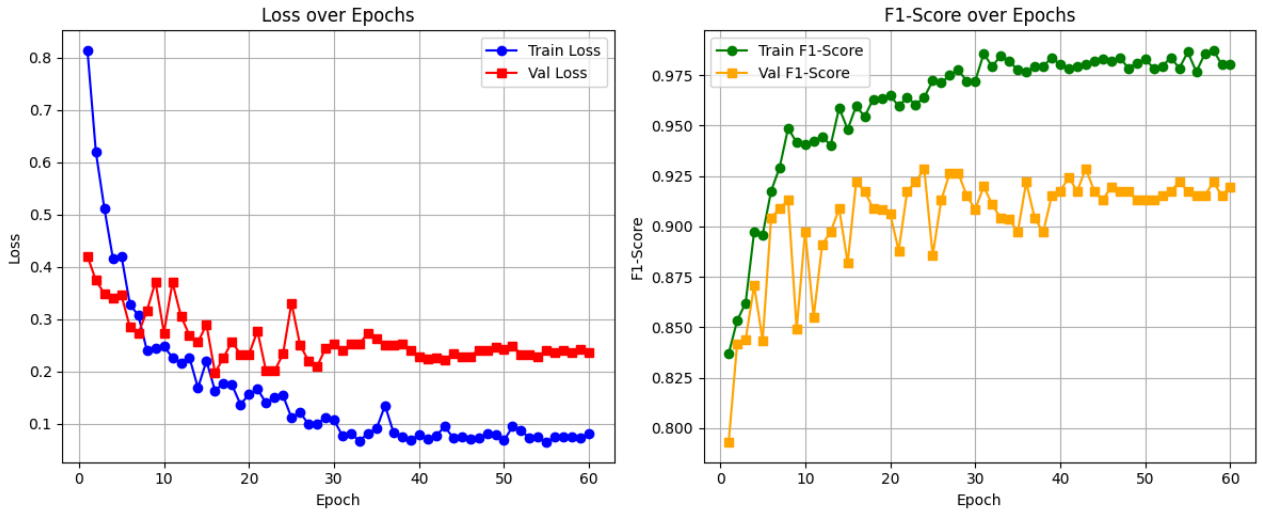
		512)						
UPFall [97]	3-layer (8, 16, 32)	9-layer (64×4, 128×3, 256×3)	0.2828	0.9288	0.2735	0.9155	0.2167	0.9265

Bảng 4.4. So sánh số khung hình mỗi giây (FPS) giữa các phương pháp tối ưu mô hình khác nhau.

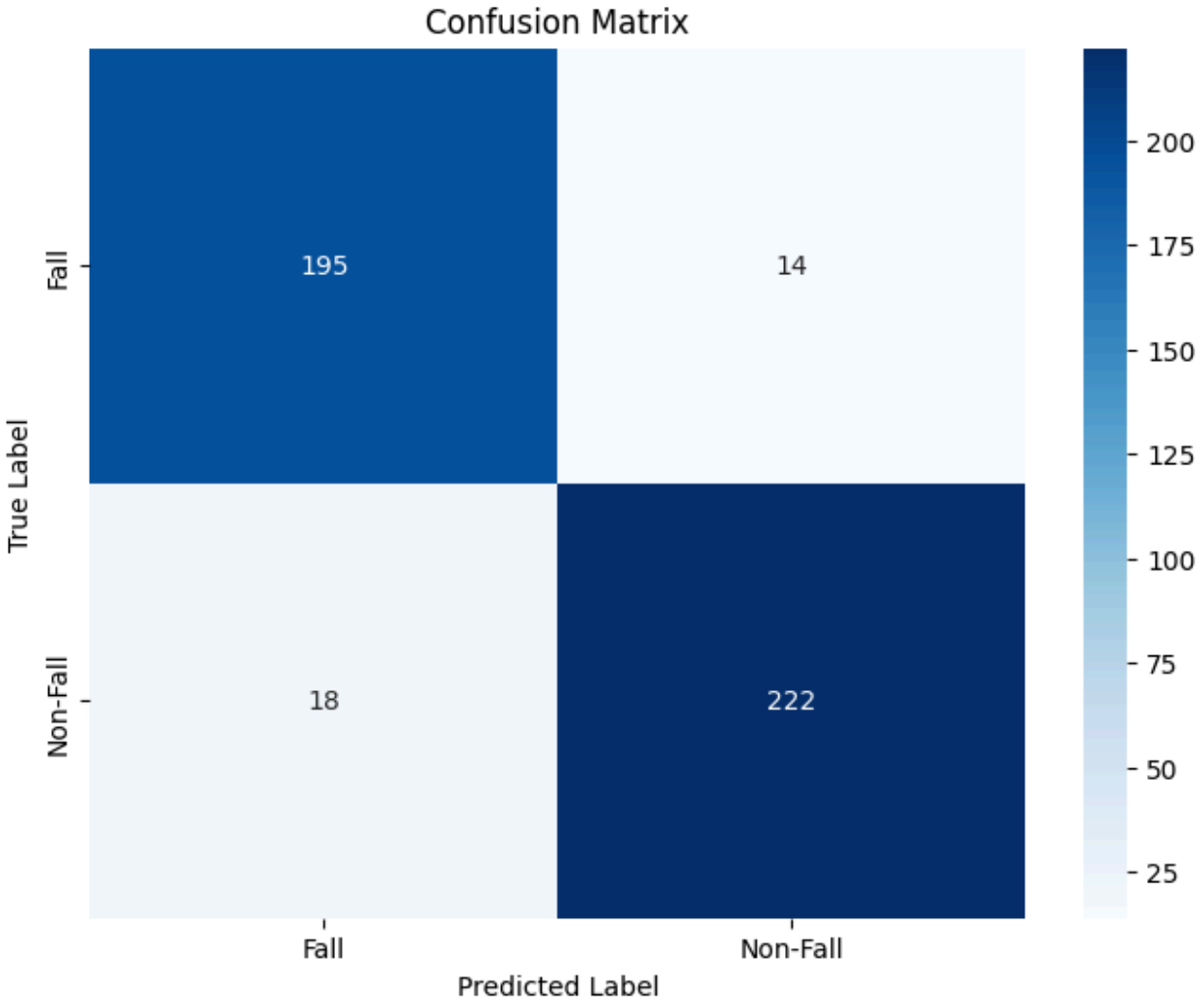
Mô hình	**FPS**
LSTM [21]	4.15
RNN [21]	3.90
MLP [21]	4.45
3D-CNN [22]	2.15
Lightweight OpenPose + RNN-LSTM [23]	10.0
Distilled ST-GCN (Được đề xuất)	**4.20**



Hình 4.1. Đồ thị hàm mất mát của mô hình sinh viên khi chưa áp dụng KD



Hình 4.2. Đồ thị hàm mất mát của mô hình sinh viên khi áp dụng KD



Hình 4.3. Kết quả ma trận nhầm lẫn

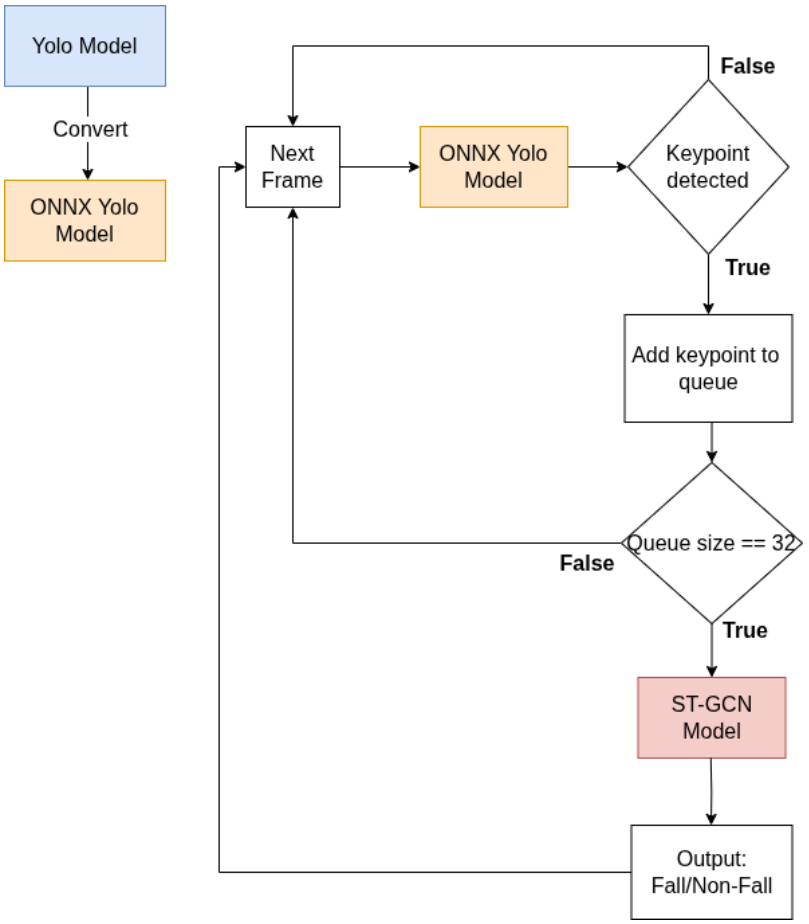
5.4. Triển khai trên thiết bị biên

Để triển khai hệ thống phát hiện té ngã người theo thời gian thực trên các nền tảng biên (edge), tôi triển khai hệ thống trên thiết bị NVIDIA Jetson Nano, một thiết bị tiêu thụ điện năng thấp, sử dụng kiến trúc ARM, được trang bị GPU Maxwell 128 lõi và RAM 4GB (xem Hình 4.1). Do bị giới hạn về tài

nguyên, cả hai mô hình YOLOv11 Pose và ST-GCN đều được chuyển đổi sang định dạng ONNX để tương thích với các công cụ suy luận được tối ưu hóa như TensorRT. Mô hình ST-GCN sau đó được lượng tử hóa xuống độ chính xác 8-bit, giúp giảm mức sử dụng bộ nhớ và tăng tốc độ tính toán mà vẫn giữ được độ chính xác. Hệ thống được hiện thực bằng ngôn ngữ Python, sử dụng OpenCV để xử lý video và ONNX Runtime để thực thi mô hình. Các điểm khớp (keypoint) được thu thập trong cửa sổ trượt 32 khung hình, và sẽ được đưa vào mô hình ST-GCN khi bộ đệm được lấp đầy.

Hình 4.4. Thiết bị biên – Bộ Kit Phát Triển NVIDIA Jetson Nano B01 – được sử dụng trong nghiên cứu.

Mặc dù đã thực hiện các tối ưu hóa, hiệu suất thời gian thực vẫn còn hạn chế. Việc suy luận bằng YOLO mất khoảng 0.2 giây cho mỗi khung hình, ngay cả khi đã tăng tốc bằng ONNX. Do đó, hệ thống phù hợp nhất cho các ứng dụng có tốc độ khung hình thấp hoặc kích hoạt theo sự kiện, chẳng hạn như phát hiện ngã hoặc phân loại cử chỉ, nơi việc bỏ qua một số khung hình là chấp nhận được. Cấu hình này cho thấy cách các pipeline học sâu cho phát hiện ngã dựa trên tư thế có thể được điều chỉnh để chạy hiệu quả trên các thiết bị biên, cho phép xây dựng các hệ thống luôn hoạt động, bảo vệ quyền riêng tư mà không cần phụ thuộc vào xử lý trên nền tảng đám mây.



Hình 4.5. Chi tiết về triển khai hệ thống trên thiết bị biên

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. KẾT QUẢ ĐẠT ĐƯỢC

Trong thời gian tìm hiểu, nghiên cứu cơ sở lý thuyết và triển khai ứng dụng công nghệ, Đồ án cũng đã đạt được những kết quả sau:

- Đóng góp mới cho cộng đồng học thuật: Là nghiên cứu đầu tiên áp dụng phương pháp KD cho tác vụ phát hiện té ngã dựa trên khung xương, mở ra hướng nghiên cứu mới kết hợp giữa nén mô hình và nhận dạng hành vi trên thiết bị biên.
- Xây dựng hệ thống phát hiện té ngã nhẹ và hiệu quả cao: Đề xuất mô hình 3-layer ST-GCN với chỉ 17,203 tham số, được huấn luyện thông qua KD từ các mô hình ST-GCN phức tạp hơn (4-layer và 9-layer), giúp giảm 200 lần số lượng tham số so với mô hình giáo viên nhưng vẫn giữ được độ chính xác cao.
- Hiệu quả KD rõ rệt: Kết quả thực nghiệm trên 5 bộ dữ liệu công khai (CaucaFall, GMDCSA-24, FallVision, URFall, và UPFall) cho thấy mô hình học trò sau khi áp dụng KD cải thiện F1-score đến 7% so với mô hình học trò ban đầu và thu hẹp đáng kể khoảng cách với mô hình giáo viên.
- Khả năng triển khai thực tế cao: Mô hình học trò đạt tốc độ xử lý trung bình 4.2 FPS trên thiết bị NVIDIA Jetson Nano, chứng minh tính khả thi trong triển khai thời gian thực trên các thiết bị biên có tài nguyên hạn chế.
- Tối ưu hoá đầy đủ cho thiết bị biên: Hệ thống được chuyển đổi sang định dạng ONNX và lượng tử hóa 8-bit, sử dụng ONNX Runtime và TensorRT để tăng tốc suy luận, đảm bảo tiêu chí hiệu quả, gọn nhẹ và dễ triển khai.
- Đánh giá toàn diện và chuẩn hoá dữ liệu: Sử dụng kỹ thuật cross-validation theo người (LOSO) và tiền xử lý khung xương bằng chuẩn hóa Min-Max và theo chiều dài cơ thể, đảm bảo mô hình học được đại diện ổn định và khả năng tổng quát tốt trên dữ liệu thực.

2. KIẾN NGHỊ VÀ HƯỚNG PHÁT TRIỂN

Kiến nghị:

- Xây dựng và cập nhật bộ dữ liệu té ngã thực tế: Đề xuất các viện dưỡng lão, bệnh viện và trung tâm chăm sóc người cao tuổi phối hợp xây dựng và chia sẻ bộ dữ liệu video té ngã trong môi trường sinh hoạt tự nhiên nhằm cải thiện tính đại diện và tính ứng dụng thực tiễn cho các mô hình AI.
- Khuyến khích mã nguồn mở và chia sẻ mô hình: Đề nghị các nhóm nghiên cứu trong nước mở rộng hợp tác, chia sẻ mô hình, pipeline xử lý và dữ liệu đã tiền xử lý (dạng khung xương) để đẩy mạnh tiến độ nghiên cứu phát hiện té ngã trên

thiết bị biên.

- Phát triển cơ chế đánh giá chuẩn hóa: Cần có bộ tiêu chuẩn đánh giá chung (theo F1-score, FPS, hiệu suất thiết bị) cho các hệ thống phát hiện té ngã để đảm bảo tính khách quan và khả năng so sánh giữa các mô hình trong nghiên cứu và triển khai thực tế.

Hướng phát triển:

- Tích hợp với hệ thống nhà thông minh và camera giám sát: Mở rộng hệ thống phát hiện té ngã để có thể tích hợp trực tiếp vào các hệ thống giám sát an ninh, camera thông minh tại gia đình hoặc bệnh viện, hỗ trợ cảnh báo khẩn cấp trong thời gian thực.
- Phát triển kiến trúc mô hình mạnh hơn nhưng vẫn nhẹ: Tiếp tục nghiên cứu các kiến trúc ST-GCN cải tiến (như 2s-AGCN, Shift-GCN) kết hợp với kỹ thuật KD để giữ được độ chính xác cao trong khi vẫn phù hợp với thiết bị biên.
- Cải tiến pipeline phát hiện nhiều đối tượng: Nâng cấp hệ thống để xử lý phát hiện té ngã cho nhiều người cùng lúc, đặc biệt trong các môi trường đông đúc như viện dưỡng lão, lớp học, trung tâm y tế.
- Triển khai học trực tuyến và học liên tục (online & continual learning): Áp dụng phương pháp học liên tục để mô hình có thể cập nhật dần dần từ dữ liệu mới trong thực tế mà không cần huấn luyện lại từ đầu, giúp mô hình thích nghi với sự thay đổi về khung cảnh, góc quay và hành vi người dùng

TÀI LIỆU THAM KHẢO

[1] World Health Organization. Fall: Fact sheets. Geneva, Switzerland: World Health Organization; 2021 [cited 2022 May 8]. Available from: https://www.who.int/news-room/fact-sheets/detail/falls.

[2] Hoang DK, Le NM, Vo-Thi UP, Nguyen HG, Ho-Pham LT, Nguyen TV. Mechanography assessment of fall risk in older adults: the Vietnam Osteoporosis Study. Journal of Cachexia, Sarcopenia and Muscle. 2021; 12(5):1161–7. Epub 2021/07/02. https://doi.org/10.1002/jcsm.12751. PMID: 34196127; PubMed Central PMCID: PMC8517351.

[3] Vu HM, Nguyen LH, Nguyen HLT, Vu GT, Nguyen CT, Hoang TN, et al. Individual and Environmental Factors Associated with Recurrent Falls in Elderly Patients Hospitalized after Falls. International Journal of Environmental Research and Public Health. 2020; 17(7). Epub 2020/04/09. https://doi.org/10.3390/ijerph17072441. PMID: 32260192; PubMed Central PMCID: PMC7177702.

[4] Muro-De-La-Herran, A.; Garcia-Zapirain, B.; Mendez-Zorrilla, A. Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications. Sensors 2014, 14, 3362–3394. [CrossRef]

[5] Rizk, H.; Yamaguchi, H.; Youssef, M.; Higashino, T. Gain without pain: Enabling fingerprinting-based indoor localization using tracking scanners. In Proceedings of the 28th International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 3–6 November 2020; pp. 550–559

[6] Lou, Z.; Wang, L.; Jiang, K.; Wei, Z.; Shen, G. Reviews of wearable healthcare systems: Materials, devices and system integration. Mater. Sci. Eng. R. Rep. 2020, 140, 100523. [CrossRef]

[7] Simon, "Quantification of human motion: gait analysisbenefits and limitations to its application to clinical problems," Journal of biomechanics, vol. 37, no. 12, pp. 1869–1880, 2004

[8] Khan, M.A.; Saboor, A.; Kim, H.c.; Park, H. A Systematic Review of Location Aware Schemes in the Internet of Things. Sensors 2021, 21, 3228. [CrossRef] Sensors 2021, 21, 5134 21 of 23

[9] Saboor, A.; Mustafa, A.; Ahmad, R.; Khan, M.A.; Haris, M.; Hameed, R. Evolution of Wireless Standards for Health Monitoring. In Proceedings of the 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON), Jaipur, India, 13–15 March 2019; pp. 268–272. [CrossRef]

[10] Ekram Alam, Abu Sufian, Paramartha Dutta, Marco Leo, Vision-based human fall detection systems using deep learning: A review, Computers in Biology and Medicine, Volume 146, 2022, 105626, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2022.105626.(https://www.sciencedirect.com/science/article/pii/S0010482522004188)

[11] N. Abbas, Y. Zhang, A. Taherkordi, and Tor Skeie, "Mobile Edge Computing: A Survey," IEEE Internet Things J., vol. 5, issue 1, pp. 450–465, 2018.

[12] Lin, B.S., Yu, T., Peng, C.W., Lin, C.H., Hsu, H.K., Lee, I.J. and Zhang, Z., 2022. Fall detection system with artificial intelligence-based edge computing. IEEE Access, 10, pp.4328-4339.

[13] Yajai, A.; Rodtook, A.; Chinnasarn, K.; Rasmequan, S.; Apichet, Y. Fall detection using directional bounding box. In Proceedings of the 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), Hatyai, Thailand, 22–24 July 2015; pp.52–57.

[14] Juang, L.H.; Wu, M.N. Fall Down Detection Under Smart Home System. J. Med. Syst. 2015, 39, 107–113. [CrossRef] [PubMed]

[15] Auvinet, E.; Rougier, C.; Meunier, J.; St-Arnaud, A.; Rousseau, J. Multiple Cameras Fall Data Set; Technical Report Number 1350; University of Montreal: Montreal, QC, Canada, 8 July 2011

[16] Chen, W., Jiang, Z., Guo, H. and Ni, X., 2020. Fall detection based on key points of human-skeleton using openpose. Symmetry, 12(5), p.744.

[17] Ramirez, Heilym, et al. ”Fall detection and activity recognition using human skeleton features.” Ieee Access 9 (2021): 33532-3354

[18] Chen, G., Duan, X., 2021. Vision-based elderly fall detection algorithm for mobile robot. In: 2021 IEEE 4th International Conference on Electronics Technology. ICET, IEEE, pp. 1197–1202.

[19] Fei, K., Wang, C., Zhang, J., Liu, Y., Xie, X. and Tu, Z., 2023. Flow-pose Net: An effective two-stream network for fall detection. The Visual Computer, 39(6), pp.2305-2320.

[20] Lin, C.B., Dong, Z., Kuan, W.K. and Huang, Y.F., 2020. A framework for fall detection based on OpenPose skeleton and LSTM/GRU models. Applied Sciences, 11(1), p.329.

[21] Pham Cuong, Nguyen Ngoc Diep, Tu Minh Phuong (2013), A Wearable Sensor based Approach to Real-Time Fall Detection and Fine-Grained Activity Recognition, Journal of Mobile Multimedia, vol. 9, no. 1&2, pp. 1526

[22] F. Bianchi, S.J. Redmond, M.R. Narayanan, S. Cerutti, N.H. Lovell (2010), Barometric pressure and triaxial accelerometry-based falls event detection, IEEE Trans. Neural Syst. Rehabil. Eng. 18 (6) 619-627

[23] Alwan M., Rajendran P.J., Kell S., Mack D., Dalal S., Wolfe M., Felder R (2006), A smart and passive floor-vibration based fall detector for elderly, In Proceedings of the 2nd Information and Communication Technologies, ICTTA ’06, Damascus, Syria; Volume 1, pp. 1003-1007.

[24] Gaya-Morey, F.X., Manresa-Yee, C. and Buades-Rubio, J.M., 2024. Deep learning for computer vision based activity recognition and fall detection of the elderly: a systematic review. Applied Intelligence, 54(19), pp.8982-9007.

[25] Nunez-Marcos, A., Azkune, G. and Arganda-Carreras, I., 2017. Vision-based fall detection with convolutional neural networks. Wireless communications and mobile computing, 2017(1), p.9474806.

[26] Nogas, J., Khan, S.S. and Mihailidis, A., 2020. Deepfall: Non-invasive fall detection with deep spatio-temporal convolutional autoencoders. Journal of Healthcare Informatics Research, 4(1), pp.50-70.

[27] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A., 2019. OpenPose: Realtime multiperson 2D pose estimation using part affinity fields.

IEEE Trans. Pattern Anal. Mach. Intell

[28] Fang, H.-S., Xie, S., Tai, Y.-W., Lu, C., 2017. RMPE: Regional multi-person pose estimation. In: ICCV

[29] Guan, Z., Li, S., Cheng, Y., Man, C., Mao, W., Wong, N., Yu, H., 2021. A video-based fall detection network by spatio-temporal joint-point model on edge devices. In: 2021 Design, Automation Test in Europe Conference Exhibition. DATE, IEEE, pp. 422–427

[30] Keskes, O. and Noumeir, R., 2021. Vision-based fall detection using st-gcn. IEEE Access, 9,pp.28224-28236.

[31] Noor, N. and Park, I.K., 2024. Factorized 3D-CNN for Real-Time Fall Detection and Action Recognition on Embedded System. IEEE Access

[32] Chang, W.J., Hsu, C.H. and Chen, L.B., 2021. A pose estimation-based fall detection methodology using artificial intelligence edge computing. IEEE Access, 9, pp.129965-129976.

[33] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

[34] M. Satyanarayanan, “The emergence of edge computing,” Computer, vol. 50, no. 1, pp. 30–39, 2017.

[35] D. Jeans. (Mar. 2019). Related’s Hudson Yards: Smart City or Surveillance City? [Online]. Available: https://therealdeal.com/2019/03/15/hudsonyards-smart-city-or-surveillance-city/

[36] M. Satyanarayanan, V. Bahl, R. Caceres, and N. Davies, “The case for VM-based cloudlets in mobile computing,” IEEE Pervasive Comput., vol. 8, no. 4, pp. 14–23, Oct./Dec. 2009.

[37] AT&T Multi-Access Edge Computing. [Online]. Available: https://www.business.att.com/products/multiaccess-edge-computing.html

[38] C.-F.-A. T. Blog. Edge Computing at Chick-Fil-A. [Online]. Available: https://medium.com/@cfatechblog/edgecomputing-at-chick-fil-a-7d67242675e2

[39] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, Deep Learning, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[40] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” IEEE Internet Things J., vol. 3, no. 5, pp. 637–646, Oct. 2016.

[41] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A survey on mobile edge computing: The communication perspective,” IEEE Commun. Surveys Tuts., vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[42] U. Drolia, K. Guo, and P. Narasimhan, “Precog: Prefetching for image recognition applications at the edge,” in Proc. ACM/IEEE Symp. Edge Comput., 2017, pp. 1–17.

[43] H. Li, K. Ota, and M. Dong, “Learning IoT in edge: Deep learning for the Internet of Things with edge computing,” IEEE Netw., vol. 32, no. 1, pp. 96–101, Jan./Feb. 2018

[44] L. N. Huynh, Y. Lee, and R. K. Balan, “DeepMon: Mobile GPU-based deep learning framework for continuous vision applications,” in Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services, 2017, pp. 82–95.

[45] Y. Kang et al., “Neurosurgeon: Collaborative intelligence between the cloud and mobile edge,” ACM SIGPLAN Notices, vol. 52, no. 4, pp. 615–629, 2017.

[46] R. Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., 2015, pp. 1310–1321.

[47] A. G. Howard et al., “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861. [Online]. Available: https://arxiv.org/abs/1704.04861

[48] W. Liu et al., “SSD: Single shot multibox detector,” in Proc. Eur. Conf. Comput. Vis. Springer, 2016, pp. 21–37.

[49] J. Redmon and A. Farhadi, “YOLO9000: Better, faster, stronger,” in Proc. IEEE CVPR, Jul. 2017, pp. 7263–7271.

[50] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <0.5 MB model size,” 2016, arXiv:1602.07360. [Online]. Available: https://arxiv.org/abs/1602.07360

[51] Tensorflow. [Online]. Available: https://www.tensorflow.org/

[52] Caffe2. [Online]. Available: https://caffe2.ai/

[53] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding,” 2015, arXiv:1510.00149. [Online].

[54] S. Han et al., “ESE: Efficient speech recognition engine with sparse LSTM on FPGA,” in Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays (FPGA), 2017, pp. 75–84

[55] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531. [Online]. Available: https://arxiv.org/abs/1503.02531

[56] S. Liu, Y. Lin, Z. Zhou, K. Nan, H. Liu, and J. Du, "DeepIoT: Compressing deep neural network structures for sensing systems with a compressor-critic framework," in Proc. SenSys, 2017, pp. 1–4

[57] L. Lai and N. Suda, "Enabling deep learning at the IoT edge," in Proc. Int. Conf. Comput.-Aided Design (ICCAD), 2018, p. 135

[58] S. Han et al., "ESE: Efficient speech recognition engine with sparse LSTM on FPGA," in Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays (FPGA), 2017, pp. 75–84

[59] S. Bhattacharya and N. D. Lane, "Sparsification and separation of deep learning layers for constrained resource inference on wearables," in Proc. 14th ACM Conf. Embedded Netw. Sensor Syst. CD-ROM (SenSys), 2016, pp. 176–189

[60] S. Yao, Y. Zhao, Z. Aston, L. Su, and T. Abdelzaher, "On-demand deep model compression for mobile devices: A usage-driven model selection

[61] N. Loc Huynh, Y. Lee, and R. K. Balan, "DeepMon: Mobile GPU-based deep learning framework for continuous vision applications," in Proc. ACM MobiSys, 2017, pp. 82–95.

[62] Edge TPU. [Online]. Available: https://cloud.google.com/edge-tpu/

[63] Z. Du et al., "Shidiannao: Shifting vision processing closer to the sensor," ACM SIGARCH Comput. Archit. News, vol. 43, no. 3, pp. 92–104, 2015.

[64] K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, and E. S. Chung. Accelerating Deep Convolutional Neural Networks Using Specialized Hardware. [Online]. Available: https://www.microsoft.com/en-us/research/wpcontent/uploads/2016/02/CNN20Whitepaper.pdf

[65] VPU. [Online]. Available: https://www.movidius.com/solutions/visionprocessing-unit

[66] S. Rivas-Gomez, A. J. Pena, D. Moloney, E. Laure, and S. Markidis, "Exploring the vision processing unit as co-processor for inference," in Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW), May 2018, pp. 589–598.

[67] Nvidia. NVIDIA EGX Edge Computing Platform. [Online]. Available: https://www.nvidia.com/enus/data-center/products/egx-edge-computing/

[68] Qualcomm. Qualcomm Neural Processing SDK for AI. [Online]. Available: https://developer.qualcomm.com/software/qualcomm-neuralprocessing-sdk

[69] M. Alzantot, Y. Wang, Z. Ren, and M. B. Srivastava, “RSTensorFlow: GPU enabled tensorflow for deep learning on commodity android devices,” in Proc. 1st Int. Workshop Deep Learn. Mobile Syst. Appl. (EMDL), 2017, pp. 7–12

[70] N. D. Lane et al., “DeepX: A software accelerator for low-power deep learning inference on mobile devices,” in Proc. 15th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN), 2016, p. 23.

[71] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, “Efficient processing of deep neural networks: A tutorial and survey,” Proc. IEEE, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[72] B. Taylor, V. S. Marco, W. Wolff, Y. Elkhatib, and Z. Wang, “Adaptive deep learning model selection on embedded systems,” in Proc. LCTES, 2018, pp. 31–43.

[73] S. Liu, Y. Lin, Z. Zhou, K. Nan, H. Liu, and J. Du, “DeepIoT: Compressing deep neural network structures for sensing systems with a compressor-critic framework,” in Proc. SenSys, 2017, pp. 1–4

[74] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[75] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF models for sequence tagging,” CoRR, vol. abs/1508.01991, pp. 1–10, Aug. 2015.

[76] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[77] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[78] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, pp. 1–14, Sep. 2014.

[79] A. Vedaldi and K. Lenc, “MatConvNet: Convolutional neural networks for MATLAB,” in Proc. 23rd ACM Int. Conf. Multimedia (MM), 2015, pp.689–692.

[80] Hinton, G., Vinyals, O. and Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

[81] Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C. and Bengio, Y., 2014. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550.

[82] Zagoruyko, S. and Komodakis, N., 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.

[83] Tung, F. and Mori, G., 2019. Similarity-preserving knowledge distillation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1365-1374).

[84] Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S. and Zhang, Z., 2019. Correlation congruence for knowledge distillation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5007-5016).

[85] Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D. and Dai, Z., 2019. Variational information distillation for knowledge transfer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9163-9171).

[86] Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D. and Dai, Z., 2019. Variational information distillation for knowledge transfer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9163-9171).

[87] Park, W., Kim, D., Lu, Y. and Cho, M., 2019. Relational knowledge distillation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3967-3976).

[88] Heo, B., Lee, M., Yun, S. and Choi, J.Y., 2019, July. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 3779-3787).

[89] Kim, J., Park, S. and Kwak, N., 2018. Paraphrasing complex network: Network compression via factor transfer. Advances in neural information processing systems, 31.

[90] Desai, C.S. and Li, G.C., 1983. A residual flow procedure and application for free surface flow in porous media. Advances in Water Resources, 6(1), pp.27-35.

[91] Huang, Z. and Wang, N., 2017. Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219.

[92] Tian, Y., Krishnan, D. and Isola, P., 2019. Contrastive representation distillation. arXiv preprint arXiv:1910.10699.

[93] Guerrero, J.C.E., España, E.M., Añasco, M.M. and Lopera, J.E.P., 2022. Dataset for human fall recognition in an uncontrolled environment. Data in brief, 45, p.108610.

[94] Alam, E., Sufian, A., Dutta, P., Leo, M. and Hameed, I.A., 2024.

GMDCSA-24: A dataset for human fall detection in videos. Data in Brief, 57, p.110892.

[95] Rahman, N.N., Mahi, A.B.S., Mistry, D., Al Masud, S.M.R., Saha, A.K., Rahman, R. and Islam, M.R., 2025. FallVision: A benchmark video dataset for fall detection. Data in Brief, 59, p.111440.

[96] Kwolek, B. and Kepski, M., 2014. Human fall detection on embedded platform using depth maps and wireless accelerometer. Computer methods and programs in biomedicine, 117(3), pp.489-501.

[97] Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J. and Peñafort-Asturiano, C., 2019. UP-fall detection dataset: A multimodal approach. Sensors, 19(9), p.1988.

[98] Gupta, S. Deep learning based human activity recognition (HAR) using wearable sensor data. Int. J. Inf. Manag. Data Insights 2021, 1, 100046. [CrossRef]

[99] Diraco, G.; Rescio, G.; Caroppo, A.; Manni, A.; Leone, A. Human Action Recognition in Smart Living Services and Applications: Context Awareness, Data Availability, Personalization, and Privacy. Sensors 2023, 23, 6040. [CrossRef] [PubMed]

[100] Shuvo, M.M.H.; Ahmed, N.; Nouduri, K.; Palaniappan, K. A Hybrid Approach for Human Activity Recognition with Support Vector Machine and 1D Convolutional Neural Network. In Proceedings of the 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 13–15 October 2020 ; pp. 1–5. [CrossRef]

[101] Rojanavasu, P.; Jantawong, P.; Jitpattanakul, A.; Mekruksavanich, S. Improving Inertial Sensor-based Human Activity Recognition using Ensemble Deep Learning. In Proceedings of the 2023 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), Phuket, Thailand, 22–25 March 2023; pp. 488–492. [CrossRef]

[102] Muhoza, A.C.; Bergeret, E.; Brdys, C.; Gary, F. Multi-Position Human Activity Recognition using a Multi-Modal Deep Convolutional Neural Network. In Proceedings of the 2023 8th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 20–23 June 2023; pp. 1–5

[103] Tao, S.; Goh, W.L.; Gao, Y. A Convolved Self-Attention Model for IMU-based Gait Detection and Human Activity Recognition. In Proceedings

of the 2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS), Hangzhou, China, 11–13 June 2023; pp. 1–5.

[104] Hassler, A.P.; Menasalvas, E.; García-García, F.J.; Rodríguez-Mañas, L.; Holzinger, A. Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. BMC Med. Inform. Decis. Mak. 2019, 19, 33. [CrossRef] [PubMed]

[103] Xu, S.; Zhang, L.; Huang, W.; Wu, H.; Song, A. Deformable convolutional networks for multimodal human activity recognition using wearable sensors. IEEE Trans. Instrum. Meas. 2022, 71, 2505414. [CrossRef]

[105] Beddiar, D.R.; Nini, B.; Sabokrou, M.; Hadid, A. Vision-based human activity recognition: A survey. Multimed. Tools Appl. 2020, 79, 30509–30555. [CrossRef]

[106] Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. IEEE Commun. Surv. Tutorials 2012, 15, 1192–1209. [CrossRef]

[107] Ke, S.R.; Thuc, H.L.U.; Lee, Y.J.; Hwang, J.N.; Yoo, J.H.; Choi, K.H. A review on video-based human activity recognition. Computers 2013, 2, 88–131. [CrossRef]

[108] Ray, A.; Kolekar, M.H.; Balasubramanian, R.; Hafiane, A. Transfer learning enhanced vision-based human activity recognition: A decade-long analysis. Int. J. Inf. Manag. Data Insights 2023, 3, 100142. [CrossRef]

[109] Kaseris, M., Kostavelis, I. and Malassiotis, S., 2024. A comprehensive survey on deep learning methods in human activity recognition. Machine Learning and Knowledge Extraction, 6(2), pp.842-876.

[110] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In NeurIPS.

[111] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully Convolutional Networks for Semantic Segmentation. In CVPR.

[112] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In NeurIPS

[113] Thoker, F.M. and Gall, J., 2019, September. Cross-modal knowledge distillation for action recognition. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 6-10). IEEE.

[114] Vu, D.Q., Le, N. and Wang, J.C., 2021. Teaching yourself: A self-knowledge distillation approach to action recognition. IEEE Access, 9,

pp.105711-105723.

[115] Garcia, N.C., Morerio, P. and Murino, V., 2018. Modality distillation with multiple stream networks for action recognition. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 103-118).

[116] Stroud, J., Ross, D., Sun, C., Deng, J. and Sukthankar, R., 2020. D3d: Distilled 3d networks for video action recognition. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 625-634).

[117] Liu, C., Jiang, Y., Du, C. and Li, Z., 2024. Enhancing action recognition from low-quality skeleton data via part-level knowledge distillation. Signal Processing, 221, p.109486.

[118] Zhang, B.; Quan, C.; Ren, F. Study on CNN in the recognition of emotion in audio and images. In Proceedings of the 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, 26–29 June 2016. [CrossRef]

[119] Pollen, D.A. Explicit neural representations, recursive neural networks and conscious visual perception. Cereb. Cortex 2003, 13, 807–814. [CrossRef] [PubMed]

[120] Using artificial neural networks to understand the human brain. Res. Featur. 2022. [CrossRef]

[121] Improvement of Neural Networks Artificial Output. Int. J. Sci. Res. (IJSR) 2017, 6, 352–361. [CrossRef]

[122] Dodia, S.; Annappa, B.; Mahesh, P.A. Recent advancements in deep learning based lung cancer detection: A systematic review. Eng. Appl. Artif. Intell. 2022, 116, 105490. [CrossRef]

[123] Ojo, M.O.; Zahid, A. Deep Learning in Controlled Environment Agriculture: A Review of Recent Advancements, Challenges and Prospects. Sensors 2022, 22, 7965. [CrossRef] [PubMed]

[124] Jarvis, R.A. A Perspective on Range Finding Techniques for Computer Vision. IEEE Trans. Pattern Anal. Mach. Intell. 1983, PAMI-5, 122–139. [CrossRef]

[125] Hussain, M.; Bird, J.; Faria, D.R. A Study on CNN Transfer Learning for Image Classification. 11 August 2018. Available online: https://research.aston.ac.uk/en/publications/a-study-on-cnn-transfer-learning-for-image-classification (accessed on 1 January 2023).

[126] Yang, R.; Yu, Y. Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis. Front. Oncol. 2021, 11, 638182. [CrossRef]
[127] Haupt, J.; Nowak, R. Compressive Sampling vs. Conventional Imaging. In Proceedings of the 2006 International Conference on Image Processing, Las Vegas, NV, USA, 26–29 June 2006; pp. 1269–1272. [CrossRef]
[128] Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. Pattern Recognit. 2018, 77, 354–377. [CrossRef]
[127] Perez, H.; Tah, J.H.M.; Mosavi, A. Deep Learning for Detecting Building Defects Using Convolutional Neural Networks. Sensors 2019, 19, 3556. [CrossRef]
[128] Hussain, M.; Al-Aqrabi, H.; Hill, R. PV-CrackNet Architecture for Filter Induced Augmentation and Micro-Cracks Detection within a Photovoltaic Manufacturing Facility. Energies 2022, 15, 8667. [CrossRef]
[129] Hussain, M.; Dhimish, M.; Holmes, V.; Mather, P. Deployment of AI-based RBF network for photovoltaics fault detection procedure. AIMS Electron. Electr. Eng. 2020, 4, 1–18. [CrossRef]
[130] Hussain, M.; Al-Aqrabi, H.; Munawar, M.; Hill, R.; Parkinson, S. Exudate Regeneration for Automated Exudate Detection in Retinal Fundus Images. IEEE Access 2022. [CrossRef]
[131] Hussain, M.; Al-Aqrabi, H.; Hill, R. Statistical Analysis and Development of an Ensemble-Based Machine Learning Model for Photovoltaic Fault Detection. Energies 2022, 15, 5492. [CrossRef]
[132] Singh, S.A.; Desai, K.A. Automated surface defect detection framework using machine vision and convolutional neural networks. J. Intell. Manuf. 2022, 34, 1995–2011. [CrossRef]
[133] Weichert, D.; Link, P.; Stoll, A.; Rüping, S.; Ihlenfeldt, S.; Wrobel, S. A review of machine learning for the optimization of production processes. Int. J. Adv. Manuf. Technol. 2019, 104, 1889–1902. [CrossRef]
[134] Wang, J.; Ma, Y.; Zhang, L.; Gao, R.X.; Wu, D. Deep learning for smart manufacturing: Methods and applications. J. Manuf. Syst. 2018, 48, 144–156. [CrossRef]
[135] Weimer, D.; Scholz-Reiter, B.; Shpitalni, M. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. CIRP Ann. 2016, 65, 417–420. [CrossRef]
[136] Kusiak, A. Smart manufacturing. Int. J. Prod. Res. 2017, 56, 508–517.

[CrossRef]

[137] Yang, J.; Li, S.; Wang, Z.; Dong, H.; Wang, J.; Tang, S. Using Deep Learning to Detect Defects in Manufacturing: A Comprehensive Survey and Current Challenges. Materials 2020, 13, 5755. [CrossRef]

[138] Soviany, P.; Ionescu, R.T. Optimizing the Trade-Off between Single-Stage and Two-Stage Deep Object Detectors using Image Difficulty Prediction. In Proceedings of the 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Timisoara, Romania, 20–23 September 2018. [CrossRef]

[139] Du, L.; Zhang, R.; Wang, X. Overview of two-stage object detection algorithms. J. Phys. Conf. Ser. 2020, 1544, 012033. [CrossRef]

[140] Sultana, F.; Sufian, A.; Dutta, P. A Review of Object Detection Models Based on Convolutional Neural Network. In Advances in Intelligent Systems and Computing; Springer: Singapore, 2020; pp. 1–16. [CrossRef]

[141] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37. [CrossRef]

[142] Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional Single Shot Detector. arXiv 2017, arXiv:1701.06659.

[143] Cheng, X.; Yu, J. RetinaNet with Difference Channel Attention and Adaptively Spatial Feature Fusion for Steel Surface Defect Detection. IEEE Trans. Instrum. Meas. 2020, 70, 2503911. [CrossRef]

[144] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]

[145] Wang, Z.J.; Turko, R.; Shaikh, O.; Park, H.; Das, N.; Hohman, F.; Kahng, M.; Chau, D.H.P. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. IEEE Trans. Vis. Comput. Graph. 2020, 27, 1396–1406. [CrossRef] [PubMed]

[146] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Commun. ACM 2017, 60, 84–90. [CrossRef]

[147] Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 2014, arXiv:1409.1556.

[148] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 12 June 2015.

[149] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 30 June 2016.

[150] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 38, 142–158. [CrossRef]

[151] Girshick, R. Fast R-CNN. In Proceedings of the International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.

[152] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. Trans. Pattern Anal. Mach. Intell. 2017, 39, 1137–1149. [CrossRef]

[153] Hussain, M., 2024. Yolov1 to v8: Unveiling each variant–a comprehensive review of yolo. IEEE access, 12, pp.42816-42833.

[154] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, ‘‘OpenPose: Realtime multi-person 2D pose estimation using part affinity fields,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[155] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, ‘‘Cascaded pyramid network for multi-person pose estimation,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7103–7112, doi: 10.1109/CVPR.2018.00742

[156] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, ‘‘Convolutional pose machines,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4724–4732, doi: 10.1109/CVPR.2016.511.

[157] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, ‘‘Learning realistic human actions from movies,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2008, pp. 1–8, doi: 10.1109/CVPR.2008.4587756.

[158] N. Dalal and B. Triggs, ‘‘Histograms of oriented gradients for human detection,’’ in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, Jun. 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177

[159] H. Bay, T. Tuytelaars, and L. Van Gool, ‘‘SURF: Speeded up robust features,’’ in Computer Vision–(ECCV) (Lecture Notes in Computer Science), vol. 3951, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Germany: Springer, 2006, doi: 10.1007/11744023_32.

[160] P. Scovanner, S. Ali, and M. Shah, ‘‘A 3-dimensional sift descriptor and its application to action recognition,’’ in Proc. 15th Int. Conf. Multimedia (MULTIMEDIA), New York, NY, USA, 2007, pp. 357–360, doi: 10.1145/1291233.1291311.

[161] H. Wang and C. Schmid, ‘‘Action recognition with improved trajectories,’’ in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 3551–3558, doi: 10.1109/ICCV.2013.441.

[162] Y. Du, W. Wang, and L. Wang, ‘‘Hierarchical recurrent neural network for skeleton based action recognition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1110–1118, doi: 10.1109/CVPR.2015.7298714.

[163] V. Veeriah, N. Zhuang, and G.-J. Qi, ‘‘Differential recurrent neural networks for action recognition,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4041–4049, doi: 10.1109/ICCV.2015.460.

[164] V. Veeriah, N. Zhuang, and G.-J. Qi, ‘‘Differential recurrent neural networks for action recognition,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4041–4049, doi: 10.1109/ICCV.2015.460.

[165] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, ‘‘Long-term recurrent convolutional networks for visual recognition and description,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 2625–2634, doi: 10.1109/CVPR.2015.7298878.

[166] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, ‘‘Beyond short snippets: Deep networks for video classification,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 4694–4702, doi: 10.1109/CVPR.2015.7299101.

[167] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, ‘‘Modeling spatialtemporal clues in a hybrid deep learning framework for video classification,’’ in Proc. 23rd ACM Int. Conf. Multimedia, New York, NY, USA, Oct. 2015, pp. 461–470, doi: 10.1145/2733373.2806222.

[168] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, ‘‘Large-scale video classification with convolutional neural networks,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1725–1732, doi: 10.1109/CVPR.2014.223.

[169] S. Ji, W. Xu, M. Yang, and K. Yu, ‘‘3D convolutional neural networks for human action recognition,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[170] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, ‘‘Deep progressive reinforcement learning for skeleton-based action recognition,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5323–5332, doi: 10.1109/CVPR.2018.00558.

[171] S. Yan, Y. Xiong, and D. Lin, ‘‘Spatial temporal graph convolutional networks for skeleton-based action recognition,’’ in Proc. AAAI Conf. Artif. Intell., 2018, vol. 32, no. 1, pp. 1–10. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/12328

[172] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, ‘‘Actionalstructural graph convolutional networks for skeleton-based action recognition,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3590–3598, doi: 10.1109/CVPR.2019.00371.

[173] X. Gao, W. Hu, J. Tang, J. Liu, and Z. Guo, ‘‘Optimized skeleton-based action recognition via sparsified graph regression,’’ in Proc. 27th ACM Int. Conf. Multimedia, New York, NY, USA, Oct. 2019, pp. 601–610, doi: 10.1145/3343031.3351170.

[174] L. Shi, Y. Zhang, J. Cheng, and H. Lu, ‘‘Two-stream adaptive graph convolutional networks for skeleton-based action recognition,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 12018–12027, doi: 10.1109/CVPR.2019.01230.

[175] W. Zheng, P. Jing, and Q. Xu, ‘‘Action recognition based on spatial temporal graph convolutional networks,’’ in Proc. 3rd Int. Conf. Comput. Sci. Appl. Eng. (CSAE), New York, NY, USA, vol. 118, 2019, pp. 1–5, doi: 10.1145/3331453.3361651

[176] K. Yang, X. Ding, and W. Chen, ‘‘Attention-based generative graph convolutional network for skeleton-based human action recognition,’’ in Proc. 3rd Int. Conf. Video Image Process., New York, NY, USA, Dec. 2019, pp. 1–6, doi: 10.1145/3376067.3376076.

[178] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, ‘‘Skeletonbased action recognition with shift graph convolutional network,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 180–189, doi: 10.1109/CVPR42600.2020.00026.

[179] Z. Huang, X. Shen, X. Tian, H. Li, J. Huang, and X.-S. Hua, ‘‘Spatiotemporal inception graph convolutional networks for skeleton-based action recognition,’’ in Proc. 28th ACM Int. Conf. Multimedia, New York, NY, USA, Oct. 2020, pp. 2122–2130, doi: 10.1145/3394171.3413666.