NHẬN DẠNG HOẠT ĐỘNG CỦA NGƯỜI TRONG VIDEO MỜ VỚI CHẮT LỌC TRI THỨC

Nguyễn Tuấn Linh*, Hoàng Anh Tú+
* Trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên
+ Nghiên cứu sinh ngành Kỹ thuật máy tính - Học viện Công nghệ BCVT

Tóm tắt: Bài báo đề xuất một kiến trúc học sâu nhận dạng hoạt động của người dựa trên hình ảnh video mờ dựa trên chắt lọc tri thức với kiến trúc thầy (teacher model) - trò (student model). Mô hình học sâu đề xuất có khả năng học và biểu diễn các đặc trưng từ video gốc và video đã được tăng cường sáng, từ đó nâng cao hiệu suất nhận dạng nhưng lại không làm tăng chi phí tính toán trong quá trình suy diễn (inference). Mô hình đề xuất tận dụng chắt lọc tri thức (knowledge distillation) trong quá trình huấn luyện trong đó mô hình thầy được huấn luyện với video đã được tăng cường sáng, trong khi mô hình trò chỉ cần huấn luyện chỉ với video gốc và các nhãn mềm (soft targets) được tạo ra bởi mô hình thầy. Các thử nghiệm cho thấy mô hình đề xuất cho kết quả cải tiến đáng kể khi so sánh với các mô hình cơ sở trên các bộ dữ liệu ARID, ARID V1.5 và Dark-481. Cụ thể, phương pháp đề xuất đạt được hiệu suất tăng tới 4,46% trên tập Dark-48 với việc chỉ sử dụng đầu vào là các video gốc, qua đó tránh được việc sử dụng hai luồng hoặc mô-đun tăng cường trong giai đoạn suy luận. Kết quả này đã chứng minh ưu điểm của việc sử dụng chắt lọc tri thức trong nhận dạng hoạt động người từ các video mờ.

Từ khóa: Nhận dạng hoạt động; Chắt lọc tri thức; Nhận dạng hoạt động trong video mờ.

I. GIỚI THIỆU

Nhận dạng hoạt động ở người luôn nhận được sự quan tâm trong lĩnh vực thị giác máy tính, được ứng dụng rộng rãi trong các hệ thống giám sát [1] và phương tiện tự hành [2, 3]. Trong những năm gần đây, ngày càng có nhiều nghiên cứu tập trung vào chủ đề này [4, 5, 6, 7, 8, 27, 28, 29, 30, 31, 32, 33]. Việc nhận dạng hoạt động trong điều kiện ánh sáng tốt đã đạt được kết quả khả quan [34, 35, 36, 37, 38, 39], tuy nhiên việc nhận dạng hoạt động trong môi trường mờ sẽ khó khăn hơn nhiều do thông tin trong video thu thập được bị suy giảm. Để giải quyết thách thức này, các nghiên cứu gần đây [9, 10, 11] đã đề xuất một số giải pháp như sử dụng phương pháp tăng cường sáng ZeroDCE và Gamma Intensity Correction (GIC) [12] để cải thiện đặc trưng và khả năng hiển thị của video, sau đó sử dụng các mạng tích chập 3D như R(2+1)D [5] hoặc 3D-ResNext [6] làm bộ phân loại backbone. Có hai cách thức để kết hợp các thành phần này bao gồm: Thứ nhất là tích hợp trực tiếp hai mô hình [10, 11]; thứ hai là sử dụng phương pháp hai luồng [5] để cải thiện độ chính xác của dự đoán hoạt động từ video mờ.

Một số nghiên cứu gần đây [10, 11] tập trung vào việc áp dụng các biện pháp tăng cường và sử dụng các video đã tăng cường làm đầu vào cho mô hình. Mặc dù các phương pháp này cải thiện các đặc trưng có trong video, tuy nhiên quá trình tăng cường cũng có thể dẫn đến mất một số đặc trưng chứa thông tin quan trọng cho nhận dạng hoạt động có trong video gốc. Một số nghiên cứu lại sử dụng phương pháp hai luồng truyền thống [13, 14] tức là lấy cả video gốc và video đã tăng cường làm đầu vào cho mô hình, cách tiếp cận này làm tăng đáng kể tài nguyên tính toán, khiến việc thực hiện dự đoán trong quá trình suy luận trở nên chậm hơn. Tuy nhiên nếu chỉ sử dụng video gốc làm đầu lại dẫn đến kết quả nhận dạng thấp hơn so với các kỹ thuật đã đề cập ở trên vì mô hình có thể nhận được ít thông tin hơn từ video gốc ban đầu nếu không có sự tăng cường.

Tóm lại, ba thách thức chính cho nghiên cứu hiện tại về nhận dạng hoạt động của con người trong video mờ là:

- Tính toàn vẹn thông tin: Đảm bảo mô hình học từ video đã tăng cường mà không làm mất thông tin cần thiết từ video gốc.
- Độ phức tạp: Tận dụng tối đa video gốc và xem xét các đặc trưng đã tăng cường mà không làm tăng độ phức tạp của mô hình.
- Hiệu suất: Cải thiện hiệu suất khi chỉ sử dụng video gốc làm đầu vào trong giai đoạn suy luận.

Theo như chúng tôi được biết, hiện nay chưa có nghiên cứu hiện có nào giải quyết được đồng thời cả ba thách thức này một cách trọn vẹn. Do đó, trong nghiên cứu này chúng tôi đã đề xuất một mô hình học đặc trưng từ cả video gốc và video đã tăng cường nhưng lại khắc phục được sự gia tăng của tài nguyên tính toán như các phương pháp hai luồng. Trong mô hình của chúng tôi, chắt lọc tri thức [15] sẽ có vai trò quan trọng giúp mô hình trò học các đặc trưng của mô hình thầy mà không cần gia tăng dữ liệu đầu vào, qua đó vừa cải thiện hiệu suất nhận dạng và hiệu quả tính toán của mô hình.

Kiến trúc của chúng tôi bao gồm một mô hình thầy, trong mô hình thầy sẽ gồm một mô-đun tăng cường sáng và một bộ phân loại hoạt động để học từ các đặc trưng đã tăng

Tác giả liên hệ: Nguyễn Tuấn Linh,
Email: ntlinh.khmt@ictu.edu.vn
Đến tòa soạn: 22/3/2025, chỉnh sửa: 26/5/2025, chấp nhận đăng: 22/6/2025.

cường, một mô hình trò chỉ chứa một bộ phân loại hoạt động để học các đặc trưng gốc và logit của thầy. Do đó, trong giai đoạn suy luận, chúng tôi chỉ cần sử dụng video gốc làm đầu vào mà không cần bất kỳ sự tăng cường bổ sung nào. Để giải quyết các thách thức về tính toàn vẹn thông tin, chúng tôi cho phép các mô hình trò học từ mô hình thầy và các video gốc, đảm bảo rằng tất cả các thông tin quan trọng đều được mô hình nắm bắt. Để giảm độ phức tạp, chúng tôi sử dụng chắt lọc tri thức thay vì cách tiếp cận hai luồng cho phép mô hình học hai loại đặc trưng mà không bao gồm chi phí bổ sung tức là chỉ cần video gốc cho mô hình trong quá trình suy luận.

Phương pháp đề xuất của chúng tôi có các đóng góp chính:

- Sử dụng cả video gốc và các đặc trưng đã tăng cường cho nhận dạng hoạt động trong điều kiện mờ.
- Chỉ sử dụng đầu vào là các video gốc mà không cần thêm đặc trưng là các video đã tăng cường trong quá trình suy luận, cải thiện được hiệu suất tính toán trong khi vẫn nâng cao được độ chính xác nhận dạng.

Tính mới trong nghiên cứu của chúng tôi nằm ở việc giải quyết đồng thời ba vấn đề chính: đảm bảo tính toàn vẹn thông tin bằng cách học từ cả video gốc và đặc trưng được tăng cường; tối ưu đánh đổi độ phức tạp bằng cách sử dụng chắt lọc tri thức thay vì phương pháp hai luồng tốn kém tài nguyên tính toán và đạt được hiệu suất nhất quán, vượt trội khi chỉ yêu cầu video gốc làm đầu vào trong suy luận.

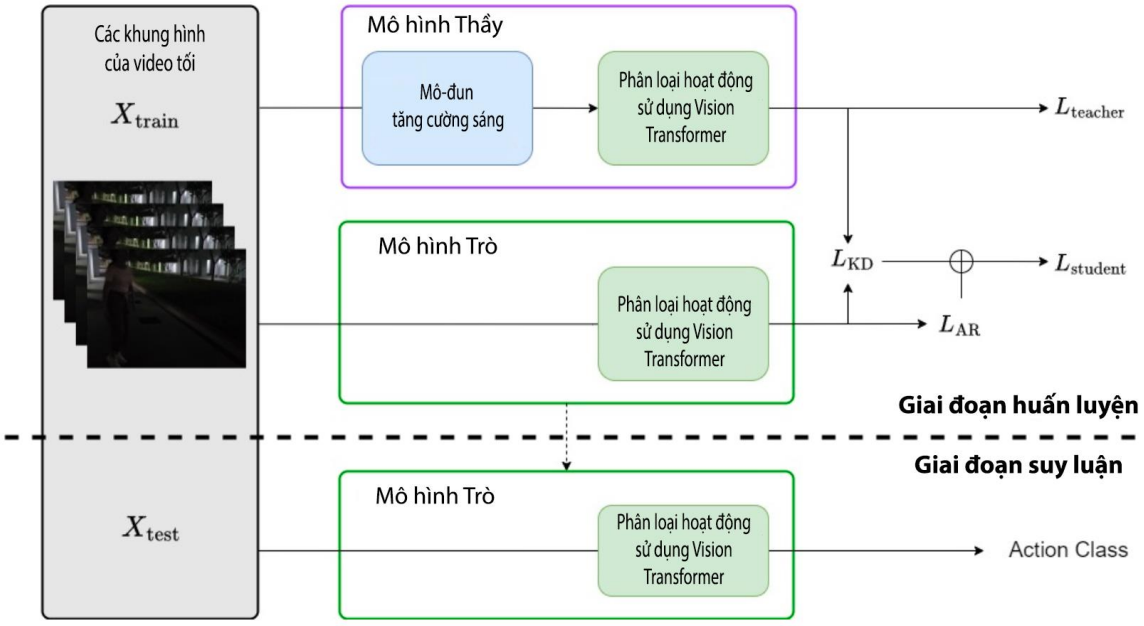
II. CÁC NGHIÊN CỨU LIÊN QUAN

A. Nhận dạng hoạt động trong video mờ

Trong những nghiên cứu gần đây, nhiều mô hình đã được đề xuất cho nhận dạng hoạt động con người, bao gồm các mô hình dựa trên 3D-CNN [4, 5, 6] và Transformers [7, 8]. Các mô hình này cho kết quả nhận dạng tốt trong điều kiện ánh sáng bình thường. Tuy nhiên, hiệu suất của chúng suy giảm khi đối mặt với video thiếu sáng, bị tối hoặc bị mờ. Do đó, các cách tiếp cận mới đã được giới thiệu để giải quyết vấn đề này, trong đó hầu hết đã chọn 3D-CNN làm backbone do tính hiệu quả của chúng. Chen và đồng sự [9] đã đề xuất DarkLight, sử dụng cả video gốc và video đã tăng cường bằng Gamma Intensity Correction (GIC) để dự đoán hoạt động. Phương pháp này đã chứng minh được tính hiệu quả của việc tăng cường ánh sáng cho nhận dạng hoạt động trong video bị tối hoặc mờ. Các thực nghiệm cũng chỉ ra rằng cách tiếp cận sử dụng hai luồng đầu vào gồm video gốc và video đã tăng cường sẽ giúp nắm bắt nhiều đặc trưng hơn so với các phương pháp ước tính luồng quang học (optical flow), do đó đạt được hiệu suất tốt hơn. Dựa trên những tiến bộ này, một số nghiên cứu [10, 11] đã cải thiện thêm độ chính xác của dự đoán hoạt động bằng cách kết hợp ZeroDCE [12] (một mô-đun tăng cường ánh sáng), các nghiên cứu này đã tích hợp ZeroDCE với các bộ phân loại backbone và đạt được hiệu suất tăng đáng kể, tuy nhiên phương pháp này lại làm gia tăng độ phức tạp tài nguyên tính toán. Do đó, mô hình đề xuất của chúng tôi sẽ giải quyết vấn đề bỏ qua tầm quan trọng của video gốc trong các nghiên cứu gần đây và cải thiện độ phức tạp tính toán do phương pháp hai luồng mang lại bằng cách áp dụng chắt lọc tri thức.

B. Chắt lọc tri thức (Knowledge Distillation)

Chắt lọc tri thức [15] đã được áp dụng cho nhiều nghiên cứu trong nhận dạng hoạt động con người, bao gồm chắt lọc tri thức đa phương thức (cross-modality) [16, 17], chắt lọc tri thức đa góc nhìn (multi-view) [18] và nhận dạng hoạt động trong các video có độ phân giải thấp hoặc bị mờ (low-resolution). Phương pháp này giúp cải thiện hiệu suất mô hình, đặc biệt khi có sự hạn chế về đầu vào trong quá trình suy luận. Lin và Tseng [18] đã đề xuất một mô hình chắt lọc tri thức đa góc nhìn cho phép mô hình học hiệu quả từ một góc nhìn trong khi lại có thể nắm bắt kiến thức từ tất cả các góc nhìn khác. Nghiên cứu này chứng minh



Hình 1. Tổng quan kiến trúc của hệ thống đề xuất

khả năng của chắt lọc tri thức trong việc cho phép các mô hình có thể học được thông tin từ nhiều nguồn dữ liệu.

Đối với việc học từ các video tăng cường ánh sáng, một số nghiên cứu [20, 21] cũng đã sử dụng chắt lọc tri thức và đạt được thành công đáng chú ý. Trong nghiên cứu của Jin và đồng sự [22] đã nhấn mạnh vai trò quan trọng của chắt lọc tri thức. Các thực nghiệm của họ cho thấy việc tích hợp chắt lọc tri thức với luồng quang học và đặc trưng RGB đã hỗ trợ hiệu quả cho quá trình huấn luyện mô hình.

Trong nghiên cứu này, chúng tôi đề xuất một cách tiếp cận mới bằng cách áp dụng trực tiếp chắt lọc tri thức vào nhiệm vụ phân loại ở tầng dưới sử dụng đặc trưng đã tăng cường. Phương pháp này cho phép mô hình trò hưởng lợi từ việc học từ video đã được xử lý tăng cường của mô hình thầy trong khi chỉ yêu cầu đầu vào là các video gốc. Bằng cách chắt lọc đặc trưng đã tăng cường cho mô hình trò, phương pháp của chúng tôi có thể tận dụng lợi thế của việc xử lý tăng cường cho video mà không cần đầu vào là các video đã tăng cường, qua đó tối ưu hóa cả hiệu suất thực thi của mô hình.

III. PHƯƠNG PHÁP ĐỀ XUẤT

A. Phát biểu bài toán

Trong các mô hình, nhận dạng hoạt động là việc cần dự đoán nhãn hoạt động từ các video đầu vào cho trước. Sử dụng phương pháp chắt lọc tri thức, chúng tôi huấn luyện một mô hình thầy sau đó sẽ chuyển kiến thức của mô hình thầy cho một mô hình trò riêng biệt. Cho $D = \{(x_i, y_i)\}_{i=0}^{n}$ là một tập dữ liệu dựa trên video, trong đó x_i là một mẫu đầu vào và y_i là lớp hoạt động. Trong giai đoạn huấn luyện, chúng tôi huấn luyện một mô hình thầy T, bao gồm một mô-đun tăng cường T_e và một bộ phân loại backbone T_c sử dụng Vision Transformer (ViT) và một mô hình trò S một cách riêng biệt.

Đối với mô hình thầy T, với các mẫu huấn luyện $X\{x_1, x_2, \dots, x_j\}$ (ở đây x_j là các khung video gốc được sử dụng cho đầu vào của mô hình) từ D_{train} làm đầu vào của T_e, mô-đun sẽ tạo ra các mẫu đã tăng cường X'. Sau đó X' sẽ được dùng làm đầu vào của T_c, nơi sẽ dự đoán xác suất của mỗi lớp trên X, ký hiệu là y^t. Hàm mất mát

$$y^t = T(x) = T_c(T_e(x)) \quad (1)$$

sau được áp dụng cho y^t và nhãn thực tế y.

Đối với mô hình trò S, với các mẫu huấn luyện $X\{x_1, x_2, \dots, x_j\}$ từ D_{train} làm đầu vào của S, mô hình tạo ra xác suất của mỗi lớp, ký hiệu là y_s. Hàm mất mát sau được áp dụng cho y^s, y^t và nhãn thực tế y.

$$y^s = S(x) \quad (2)$$

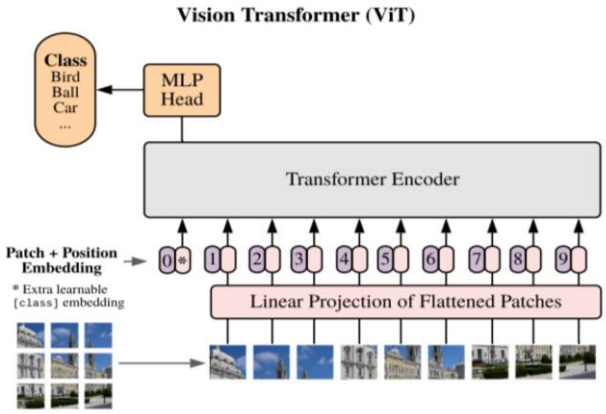
Trong giai đoạn suy luận, chỉ mô hình trò S sẽ được sử dụng để dự đoán.

B. Kiến trúc Tổng thể

Kiến trúc tổng thể được trình bày trong hình 1, kiến trúc bao gồm hai thành phần chính là mô hình thầy và mô hình trò. Mô hình thầy bao gồm một mô-đun được sử dụng để tăng cường video gốc, tiếp đến là một bộ phân loại hoạt động sử dụng Vision Transformer được mô tả chi tiết trong hình 2. Trong mô hình thầy, trước tiên cần thực hiện tăng cường các khung hình trong video gốc để cải thiện khả năng hiển thị nhằm trích xuất các đặc trưng có thể bị mất trong video mờ hoặc bị tối. Các đặc trưng đã tăng cường này sau đó được đưa vào bộ phân loại hoạt động sử dụng Vision Transformer (ViT) để tạo ra các biểu diễn đặc trưng mới.

ViT áp dụng một kiến trúc giống Transformer trên các "phần vá" (patches) của hình ảnh. Bằng cách chia hình ảnh thành các phần vá cố định, sau đó nhúng tuyến tính và thêm các nhúng vị trí, chuỗi vector kết quả được đưa vào một bộ mã hóa Transformer tiêu chuẩn. Cách tiếp cận này giúp ViT có khả năng nắm bắt các mối quan hệ khoảng cách dài và các đặc trưng phức tạp vốn là ưu điểm nổi bật của kiến trúc Transformer so với các mạng truyền thống hơn như CNN, đặc biệt hữu ích trong các video mờ nơi thông tin cục bộ có thể bị suy giảm.

Khi ViT được sử dụng trong mô hình Trò của kiến trúc chắt lọc tri thức giúp giải quyết thách thức về độ phức tạp. Mô hình có thể học các đặc trưng từ cả video gốc và các đặc trưng từ video đã tăng cường sáng (thông qua "nhãn mềm" của mô hình Thầy) mà không làm tăng chi phí tính toán đáng kể trong quá trình suy luận, đây là một mục tiêu quan trọng của mô hình đề xuất.



Hình 2. Mô hình phân loại khung hình Vision Transformer [40]

Sau khi huấn luyện mô hình thầy, chúng tôi huấn luyện mô hình trò bằng cách lấy các khung hình video gốc không tăng cường làm đầu vào. Cách tiếp cận này cho phép mô hình trò học trực tiếp từ dữ liệu gốc và đảm bảo rằng mô hình không phụ thuộc vào đầu vào là các video đã tăng cường trong quá trình suy luận. Mô hình trò là một bộ phân loại hoạt động, quá trình huấn luyện mô hình trò là một quá trình học gồm 2 giai đoạn:

- Học trực tiếp (Direct Learning): Mô hình trò học trực tiếp từ các nhãn thực tế (ground truth labels).
- Học chắt lọc tri thức (Distillation Learning): Mô hình trò được huấn luyện sử dụng các logit từ mô hình thầy làm mục tiêu mềm (soft targets), cho phép mô hình trò có khả năng học các đặc trưng đã được trích xuất từ các khung hình video đã tăng cường bởi mô hình thầy.

Kiến trúc đề xuất của chúng tôi tận dụng được cả các đặc trưng từ video đã tăng cường và các đặc trưng từ video gốc, qua đó tối ưu hóa hiệu suất của mô hình và không làm gia tăng chi phí tính toán do chỉ cần sử dụng video gốc trong giai đoạn suy luận.

C. Chắt lọc tri thức (Knowledge Distillation)

Trích chọn các đặc trưng từ video đã tăng cường: Quá trình huấn luyện mô hình thầy bắt đầu bằng việc tiến hành xử lý tăng cường chất lượng hoặc độ sáng cho các khung hình của video gốc. Mô-đun tăng cường biến đổi đầu vào I thành I'. Việc tăng cường này có thể cải thiện khả năng hiển thị của các khung hình, giúp cho việc nhận dạng hoạt động trong video mờ hoặc tối trở lên tốt hơn. Sau khi tăng cường, các khung hình đã tăng cường I' được đưa vào bộ phân loại hoạt động để dự đoán. Bộ phân loại hoạt động sẽ xử lý các khung hình đầu vào thành các logit giúp nắm bắt các thông tin quan trọng của các nhãn hoạt động. Hàm mất mát Cross-Entropy chuẩn được áp dụng để huấn luyện mô hình thầy:

$$L_{teacher} = CrossEntropy(y, y^t) \quad (3)$$

trong đó y là nhãn lớp của video đầu vào và y^t là logit được trích xuất bởi mô hình thầy.

Học các biểu diễn gốc: Không giống như mô hình thầy, mô hình trò trực tiếp lấy các khung hình từ video gốc I làm đầu vào. Bộ phân loại được sử dụng trên mô hình trò giống với bộ phân loại của mô hình thầy, tuy nhiên mô hình trò không cần sử dụng mô-đun tăng cường. Đầu vào được xử lý bởi mô hình trò sẽ tạo ra kết quả y^s, một hàm mất mát cross-entropy được áp dụng để hiệu chỉnh dự đoán của mô hình trò với nhãn thực tế.

$$L_{AR} = CrossEntropy(y, y^s) \quad (4)$$

Học kép (Dual Knowledge Learning): Ngoài việc học từ nhãn thực tế, một quá trình chắt lọc tri thức được áp dụng cho mô hình trò vì nó có thể học từ logit y^t được tạo ra bởi mô hình thầy. Điều này được thực hiện bằng cách giảm thiểu sự phân kỳ Kullback-Leibler (KL divergence) giữa đầu ra y^s của mô hình trò và các mục tiêu mềm do mô hình thầy cung cấp, từ đó sẽ gián tiếp chuyển kiến thức từ các đặc trưng đã tăng cường từ mô hình thầy sang mô hình trò.

$$L_{KD} = KL(y^t, y^s) \quad (5)$$

Tổng hàm mất mát để huấn luyện mô hình trò là sự kết hợp của hai hàm mất mát gồm hàm mất mát Cross-Entropy để học từ nhãn thực tế và phân kỳ KL để truyền kiến thức từ mô hình thầy sang mô hình trò. Tổng hàm mất mát cho mô hình trò được ký hiệu là:

$$L_{student} = \alpha L_{AR} + \beta L_{KD} \quad (6)$$

trong đó α và β là trọng số của các hàm mất mát riêng biệt để cân bằng tầm quan trọng của hai nguồn huấn luyện.

I. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Phần này trình bày về thử nghiệm để đánh giá phương pháp phát hiện vận động bất thường đã trình bày ở trên.

IV. TẬP DỮ LIỆU THỬ NGHIỆM

A. Các cài đặt thử nghiệm

Tập dữ liệu: Chúng tôi đã đánh giá mô hình đề xuất trên ba tập dữ liệu bao gồm: ARID, ARID V1.5 [23] và Dark-48 [11]. ARID [23] là một tập dữ liệu chuẩn cho nhận dạng hoạt động con người trong điều kiện video bị mờ và tối chứa hơn 3780 video với 11 lớp hoạt động. Tập dữ liệu ARID V1.5 [23] với số lượng video mờ và tối được mở rộng lên đến 5572 và 24 lớp hoạt động. Tập dữ liệu Dark-48 [11] bao gồm 8815 video tối, 48 lớp hoạt động với hơn 100 video mỗi lớp. Tập dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 8:2. Các cài đặt huấn luyện và kiểm tra trong thử nghiệm được tham khảo trong nghiên cứu [11, 23].

Chi tiết các cài đặt thử nghiệm: So sánh với thử nghiệm từ [9], chúng tôi đã chọn ViT với nhiệm vụ phía sau là mô hình BERT [24] thay thế cho cách làm truyền thống là sử dụng pooling trung bình toàn cục theo thời gian làm bộ phân loại cho mô hình backbone của chúng tôi. Mô hình backbone được huấn luyện trước (pre-trained) trên IG65M [25]. Đối với mô-đun tăng cường sáng của mô hình thầy, chúng tôi đã chọn ZeroDCE được sử dụng trong nghiên cứu [12] được thực hiện trên các khung hình của video gốc. Sau đó các khung hình được thay đổi kích thước thành 112 x 112 pixel, như vậy đầu vào cuối cùng sẽ là 3 x 64 x 112 x 112 với kích thước batch là 2. Chúng tôi đã huấn luyện mô hình thầy và mô hình trò với trình tối ưu hóa AdamW [26], tốc độ học là 0.0001. Các tham số cho hàm mất mát của mô hình trò được đặt là 1 cho cả α và β.

Ma trận (Metric): Đối với tác vụ này, chúng tôi đã ghi lại độ chính xác top-1 và top-5 để đánh giá hiệu suất của mô hình. Do tập dữ liệu ARID [23] chỉ chứa 11 lớp hoạt động và hầu hết các công trình trước đây đã gần đạt 100% độ chính xác top-5 do đó chúng tôi chủ yếu đánh giá độ chính xác top-1 cho ARID và ARID V1.5.

B. Đánh giá hiệu của mô hình

Trong phần này, chúng tôi tập trung vào đánh giá hiệu suất của mô hình khi sử dụng tập dữ liệu ARID. Chúng tôi tiến hành so sánh kết quả của mô hình đã được huấn luyện với phương pháp đề xuất và không sử dụng phương pháp đề xuất. Ngoài ra, chúng tôi còn tiến hành sánh hiệu suất giữa mô hình thầy và mô hình trò và thấy rằng mô hình trò có thể đạt được kết quả tốt hơn ngay cả khi không sử dụng các khung hình đã tăng cường sáng. Bảng 1 cung cấp kết quả chi tiết về hiệu suất của mô hình thầy và mô hình trò có sử dụng phương pháp đề xuất và hiệu suất của kiến trúc tương tự nhưng không sử dụng phương pháp đề xuất.

Bảng 1: So sánh Hiệu suất của Mô hình đề xuất trên tập dữ liệu ARID

Mô hình	**Độ chính xác Top-1 (%)**
R(2+1)D + BERT [9]	92,44
Teacher: ZeroDCE + ViT + BERT	95,87
Student: ViT + BERT	**97,36**

Như trong Bảng 1, phương pháp huấn luyện sử dụng chắt lọc tri thức của chúng tôi đã cải thiện hiệu suất tăng

thêm 4,92%. Với kiến thức bổ sung do mô hình thầy cung cấp, mô hình trò có thể tận dụng lợi thế của biểu diễn đã tăng cường ngay cả khi không có đầu vào là các đặc trưng đã tăng.

Mặc dù mô hình thầy hoạt động trên dữ liệu đã được xử lý tăng cường sáng (được thiết kế để cung cấp nhiều thông tin hơn trong điều kiện video bị mờ, tối), mô hình Trò vẫn đạt được độ chính xác vượt trội hơn (97,36% trên tập dữ liệu ARID so với 95,87% của mô hình thầy). Sự cải thiện 1,49% này đến từ chiến lược học tập độc đáo của mô hình đề xuất của chúng tôi:

Mô hình thầy do được huấn luyện trên video đã tăng cường nên có khả năng trích xuất các đặc trưng nâng cao và tạo ra các "soft target" (logits) chứa đựng thông tin sâu sắc về phân loại hành động. Nhờ quá trình chưng cất thi thức này cho phép mô hình trò học hỏi từ các "soft target" được tạo ra từ mô hình thầy, về cơ bản là tiếp thu kiến thức từ các đặc trưng đã được tăng cường mà không cần trực tiếp xử lý video tăng cường trong giai đoạn suy luận. Từ đó giúp mô hình trò tận dụng được lợi ích của việc tăng cường ánh sáng mà không làm tăng chi phí tính toán khi đưa vào thực tế1....

Mặc dù các phương pháp tăng cường ánh sáng cải thiện khả năng hiển thị của video tối, nhưng quá trình này thường dẫn đến việc mất mát thông tin quan trọng trong video gốc (chưa được tăng cường). Do đó mô hình trò được đào tạo trực tiếp trên video gốc nên cho phép nó giữ lại và khai thác thông tin cốt lõi vốn có trong dữ liệu ban đầu. Qua đó khắc phục hạn chế của các nghiên cứu gần đây thường bỏ qua tầm quan trọng của video gốc.

Bằng cách kết hợp tri thức được chưng cất từ các đặc trưng tăng cường (từ mô hình thầy) với thông tin quan trọng được bảo toàn từ video gốc, mô hình trò đạt được sự hiểu biết toàn diện hơn về hoạt động có trong video. Nó có thể truy cập thông tin bổ sung từ các đặc trưng đã được tăng cường mà không cần đầu vào tăng cường trực tiếp. Sự kết hợp này cho phép mô hình trò vượt qua các hạn chế của việc chỉ dựa vào một loại đầu vào (ví dụ, video gốc thiếu thông tin trong môi trường tối) hoặc chỉ dựa vào video tăng cường (có thể bị mất thông tin gốc).

So sánh giữa mô hình trò và mô hình thầy cho thấy rằng ngay cả khi mô hình trò sử dụng kiến trúc đơn giản hơn với việc không sử dụng các khung hình đã tăng cường nhưng nó vẫn đạt được hiệu suất tăng so với mô hình thầy. Điều này chỉ ra rằng ngoài kiến thức được chắt lọc từ các đặc trưng đã tăng cường, video gốc cũng chứa thông tin quan trọng giúp cải thiện hiệu suất mô hình. Bằng cách học từ đầu vào là các video gốc, mô hình trò đã có được các thông tin bổ sung cho các đặc trưng đã tăng cường, dẫn đến hiệu suất tốt hơn mô hình thầy.

C. So sánh với các phương pháp nổi bật đã công bố gần đây

Các nghiên cứu trước đây đã cho thấy hầu hết các phương pháp trên tập dữ liệu ARID đã đạt gần 100% độ chính xác Top-5, cho thấy đây là một tập dữ liệu tương đối "dễ" với số lượng lớp ít. Khi một mô hình đã đạt độ chính xác rất cao, "không gian" để cải thiện thêm sẽ bị hạn chế. Do đó, việc mô hình trò vẫn có thể vượt qua mô hình thầy trên ARID là một minh chứng cho hiệu quả của nó, ngay cả khi các mô hình khác cũng đã có điểm số cao. Từ đó có thể thấy mô hình trò có khả năng tinh chỉnh và tối ưu hóa thông tin học được để đạt đến giới hạn trên của tập dữ liệu này.

Chúng tôi tiến hành các thực nghiệm khác để so sánh phương pháp của chúng tôi với các phương pháp nổi bật đã công bố gần đây về nhận dạng hành động người trong điều kiện bị mờ hoặc tối như DarkLight [9], DTCM [11] và R(2+1)D-GCN+BERT [10]. Kết quả từ Bảng 2 và Bảng 3 cho thấy phương pháp đề xuất của chúng tôi đạt được kết quả tốt nhất. Bảng 4 chỉ ra rằng mô hình của chúng tôi đạt độ chính xác Top-1 là 51,14% trên tập dữ liệu Dark-48. Kết quả này có sự cải thiện đáng kể lên đến 4,46% so với kết quả tốt nhất được công bố trước đó trên Dark-48. Những kết quả này một lần nữa chỉ ra rằng sử dụng chắt lọc tri thức đã cho phép mô hình đạt được hiệu suất tốt nhất trong khi chỉ sử dụng đầu vào video gốc trong quá trình kiểm tra.

Bảng 2: So sánh kết quả trên tập dữ liệu ARID

Mô hình	**Độ chính xác Top-1 (%)**
I3D-RGB	68,29
I3D Two-stream	72,78
3D-ResNext-101	74,73
DarkLight	94,04
DTCM	96,36
R(2+1)D-GCN+BERT	96,60
Mô hình của chúng tôi	**97,36**

Bảng 3: So sánh kết quả trên ARID V1.5

Mô hình	**Độ chính xác Top-1 (%)**
I3D-RGB	48,75
I3D Two-stream	51,24
DarkLight	84,13
R(2+1)D-GCN+BERT	86,93
Mô hình của chúng tôi	**88,15**

Bảng 4: So sánh kết quả trên tập dữ liệu Dark-48

Mô hình	**Độ chính xác Top-1 (%)**	**Độ chính xác Top-5 (%)**
I3D-RGB	32.25	65.35
3D-ResNext-101	37.23	68.86
DarkLight	42.27	70.47
DTCM	46.68	75.92
Mô hình của chúng tôi	**51.14**	**78.49**

Mô hình đề xuất của chúng tôi đạt độ chính xác Top-1 là 51,14% và Top-5 là 78,49% trên Dark-48. Điều này thể hiện sự cải thiện đáng kể nhất về mặt tuyệt đối (4,46% độ chính xác Top-1) so với các phương pháp công bố trước đây.

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đề xuất một mô hình cho nhận dạng hoạt động người trong điều kiện video bị mờ hoặc tối, nhấn mạnh tầm quan trọng của việc sử dụng cả video gốc và đặc trưng đã tăng cường để ngăn chặn việc mất thông tin gốc. Hơn nữa mô hình được đề xuất đã tránh được việc gia tăng chi phí tính toán khi sử dụng phương

pháp hai luồng vào. Chúng tôi đã chưng cất hiệu quả tri thức về tăng cường ánh sáng cho mô hình trò, cho phép mô hình trò chỉ sử dụng video gốc làm đầu vào trong quá trình suy luận nhưng lại đạt được kết quả tốt hơn mô hình thầy. Việc so sánh với các công bố nổi bật trên tập dữ liệu ARID và Dark-48 một lần nữa chứng minh tính hiệu quả của phương pháp đề xuất của chúng tôi. Trong tương lai chúng tôi có thể tiếp tục nghiên cứu tinh chỉnh kiến trúc để có thể có kết quả nhận dạng tốt hơn với các video có nhiễu phức tạp hoặc các đối tượng trong video chuyển động với tốt độ nhanh, độ phân giải thấp.

TÀI LIỆU THAM KHẢO

[1] Jindong, W. , Yiqiang, C., Shuji, H., Xiaohui, P., Lisha, H.: Deep learning for sensor-based activity recognition: A survey**.** Pattern Recognition Letters 119: 3-11 (2019)

[2] Pham, C., Nguyen, N-D., Tu, M-P.; e-Shoes: Smart Shoes for Unobtrusive Human Activity Recognition. In proc. of 9th IEEE International Conference on Knowledge Systems Engineering (IEEE KSE) 2017. 269-274

[3] Nguyen, N., D. Pham, C., Tu, M., P.; Motion Primitive Forests for Human Activity Recognition Using Wearable Sensors. In proc. of the 14th Pacific Rim International Conference on Artificial Intelligence (PRICAI) 2016. 340-353

[4] Jie, J., Qiang, Y. , Jeffrey, J. P.: Sensor-Based Abnormal Human-Activity Detection. IEEE Trans. Knowl. Data Eng. 20(8): 1082-1090 (2008)

[5] Tran, T-H., Le, T-L., Pham, D-T., Hoang, V-N., Khong, V-M., Tran, Q-T, Nguyen, T-S., Pham, C.; A Multi-modal Multi-view Dataset for Human Fall Analysis and Preliminary Investigation on Modality. In the proc. the 24th International Conference on Pattern Recognition (ICPR), 1947-1952. Bejing, China, 2018

[6] Y. Yao, F. Wang, J. Wang, and D.D. Zeng, “Rule þ Exception Strategies for Security Information Analysis,” IEEE Intelligent Systems, vol. 20, no. 5, pp. 52-57, Sept./Oct. 2005.

[7] P. Jarvis, T.F. Lunt, and K.L. Myers, “Identifying Terrorist Activity with AI Plan Recognition Technology,” Proc. 19th Nat’l Conf. Artificial Intelligence (AAAI ’04), pp. 858-863, July 2004.

[8] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford, “A Hybrid Discriminative/Generative Approach for Modeling Human Activities,” Proc. 19th Int’l Joint Conf. Articial Intelligence (IJCAI ’05), pp. 766-772, July-Aug. 2005.

[9] D.J. Patterson, L. Liao, L. Fox, and H. Kautz, “Inferring High-Level Behavior from Low-Level Sensors,” Proc. Fifth Int’l Conf. Ubiquitous Computing (UbiComp ’03), pp. 73-89, Oct. 2003.

[10] B. Scho¨lkopf, J. Platt, J. Shawe-Taylor, and A. Smola, “Estimating the Support of a High-Dimensional Distribution,” Neural Computation, vol. 13, no. 7, pp. 1443-1471, July 2001.

[11] L. Liao, D. Fox, and H. Kautz, “Learning and Inferring Transportation Routines,” Proc. 19th Nat’l Conf. Artificial Intelligence (AAAI ’04), pp. 348-353, July 2004.

[12] J. Yin, X. Chai, and Q. Yang, “High-Level Goal Recognition in a Wireless LAN,” Proc. 19th Nat’l Conf. in Artificial Intelligence (AAAI ’04), pp. 578-584, July 2004.

[13] P. Lukowicz, J. Ward, H. Junker, M. Sta¨ger, G. Tro¨ ster, A. Atrash, and T. Starner, “Recognizing Workshop Activity Using Body Worn Microphones and Accelerometers,” Proc. Second Int’l Conf. Pervasive Computing (Pervasive ’04), pp. 18-32, Apr. 2004.

[14] T. Xiang and S. Gong, “Video Behaviour Profiling and Abnormality Detection without Manual Labeling,” Proc. IEEE Int’l Conf. Computer Vision (ICCV ’05), pp. 1238-1245, Oct. 2005.

[15] T. Duong, H. Bui, D. Phung, and S. Venkatesh, “Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition (CVPR ’05), pp. 838-845, June 2005.

[16] S.D. Bay and M. Schwabacher, “Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule,” Proc. Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’03), pp. 29-38, Aug. 2003.

[17] C. Elkan, “The Foundations of Cost-Sensitive Learning,” Proc. 17th Int’l Joint Conf. Articial Intelligence (IJCAI ’01), pp. 973-978, Aug. 2001.

[18] J. Ma and S. Perkins, “Time-Series Novelty Detection Using One- Class Support Vector Machines,” Proc. Int’l Joint Conf. Neural Networks (IJCNN ’03), pp. 1741-1745, July 2003.

[19] M.M. Breunig, H.P. Kriegel, R. Ng, and J. Sander, “Identifying Density-Based Local Outliers,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’00), pp. 93-104, May 2000.

[20] K.M. Ting, “A Comparative Study of Cost-Sensitive Boosting Algorithms,” Proc. 17th Int’l Conf. Machine Learning (ICML ’00), pp. 983-990, June-July 2000.

[21] A.P. Bradley, “The Use of the Area under the ROC Curve in the Evaluation of Machine Learing Algorithms,” Pattern Recognition, vol. 30, pp. 1145-1159, 1997.

[22] C.X. Ling, V.S. Sheng, and Q. Yang, “Test Strategies for Cost-Sensitive Decision Trees,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 8, pp. 1055-1067, Aug. 2006.

[23] Q. Yang, C. Ling, X. Chai, and R. Pan, “Test-Cost Sensitive Classification on Data with Missing Values,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 5, pp. 626-638, May 2006.

[24] P. Domingos, “Metacost: A General Method for Making Classifiers Cost-Sensitive,” Proc. Fifth Int’l Conf. Knowledge Discovery and Data Mining (KDD ’99), pp. 155-164, Aug. 1999.

[25] U. Knoll, G. Nakhaeizadeh, and B. Tausend, “Cost-Sensitive Pruning of Decision Trees,” Proc. 18th European Conf. Machine Learning (ECML ’94), pp. 383-386, Apr. 1994.

[26] M. Kukar and I. Kononenko, “Cost-Sensitive Learning with Neural Networks,” Proc. 13th European Conf. Artificial Intelligence (ECAI ’98), pp. 445-449, Aug. 1998.

[27] G. Fumera and F. Roli, “Cost-Sensitive Learning in Support Vector Machines,” Proc. Workshop Machine Learning, Methods and Applications, held in the Context of the Eighth Meeting of the Italian Assoc. Of Artificial Intelligence (AI*IA ’02), Sept. 2002.

[28] P. Chan and S. Stolfo, “Toward Scalable Learning with Non-Uniform Class and Cost Distributions,” Proc. Fourth Int’l Conf. Knowledge Discovery and Data Mining (KDD ’98), pp. 164-168, Aug. 1998.

[29] Geoffrey McLachlan and Thriyambakam Krishnan. The EM Algorithm and Extensions. John Wiley & Sons, New York, 1996. [23] B. Scho¨lkopf, J. Platt, J. Shawe-Taylor, and A. Smola, “Estimating the Support of a High-Dimensional Distribution,” Neural Computation, vol. 13, no. 7, pp. 1443-1471, July 2001.

[30] B. Scho¨lkopf, J. Platt, J. Shawe-Taylor, and A. Smola, "Estimating the Support of a High-Dimensional Distribution," Neural Computation, vol. 13, no. 7, pp. 1443-1471, July 2001.

[31] I.W. Tsang, J.T. Kwok, B. Mak, K. Zhang, and J.J. Pan, "Fast Speaker Adaptation via Maximum Pernalized Likelihood Kernel Regression," Proc. Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP '06), May 2006.

[32] Pham, C., Nguyen, T.; Real-Time Traffic Activity Detection Using Mobile Devices. In proc. of the 10th ACM International Conference on Ubiquitous Information Management and Communications (ACM IMCOM) 2016. 641-647

[33] Pham, C.; MobiRAR: Real-Time Human Activity Recognition Using Mobile Devices. In proc. of the 7th IEEE International Conference on Knowledge Systems Engineering (IEEE KSE) 2015. 144-149

[34] Nguyen, N., D., Pham, C., Tu, M.,P.; A Classifier Approach to Real-Time Fall Detection Using Low-Cost Wearable Sensors. In Proc. of the 5th International Symposium on Information and Communication Technology (SoICT) 2014. 14-20

[35] Pham, C.; MobiCough: Real-Time Cough Detection and Monitoring Using Low-Cost Mobile Devices. In proc. of the 8th Asean Conference on Intelligent Information and database systems (ACIIDS) 2016. 300-309

[36] Visalakshmi, S., Paul, E., Watson, P., Pham, C., Jackson, D., Olivier, P. 2011; Distributed Event Processing for Activity Recognition. In the Proceedings of the 5th ACM International Conference on Distributed Event-Based Systems (ACM DEBS) 2011 (New York, NY, 11-14 July 2011). 371-372

[37] Nguyen, L., Le, A., T., Pham, C.; The Internet-of-Things based Fall Detection Using Fusion Feature. Accepted at the 10th IEEE International Conference on Knowledge Systems Engineering (IEEE KSE) 2018. 129-134.

[38] Duc Duy Nguyen, Lam Thanh Nguyen, Yifeng Huang, Cuong Pham, Minh Hoai "Class-Agnostic Repetitive Action Counting Using Wearable Devices". IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 47 (6). 4984 - 4995.

[39] Duc-Anh Nguyen, Cuong Pham, Nhien-An Le-Khac "Virtual fusion with contrastive learning for single sensor-based activity recognition". IEEE Sensors journal (2024) 24 (15) 25041 - 25048.

[39] Yifeng Huang, Duc Duy Nguyen, Lam Nguyen, Minh Hoai, Cuong Pham "Count what you want: exemplar identification and few-shot counting of human actions in the wild". Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2024, 38 (9) 10057-10065.

[40] Liu, Ze; Lin, Yutong; Cao, Yue; Hu, Han; Wei, Yixuan; Zhang, Zheng; Lin, Stephen; Guo, Baining (2021-03-25). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". arXiv:2103.14030.

HUMAN ACTION RECOGNITION IN DARK VIDEOS WITH KNOWLEDGE DISTILLATION

Abstract: The paper proposes a deep learning architecture for human action recognition in dark videos based on knowledge distillation with a teacher-student framework. The proposed deep learning model is capable of learning and representing features from both original and enhanced videos, thereby improving recognition performance without introducing additional computational cost during the inference phase. The proposed model utilizes knowledge distillation during training, where the teacher model is trained with enhanced video, while the student model is trained solely with original video and soft targets generated by the teacher model. Experiments show that the proposed model yields significant improvements when compared to baseline models on the ARID, ARID V1.5, and Dark-48 datasets. Specifically, the proposed method achieved a performance improvement of up to 4.46% on the Dark-48 dataset using only original video inputs, thereby avoiding the use of two-stream frameworks or enhancement modules during inference. This result demonstrates the advantages of using knowledge distillation for human action recognition in dark videos.

Keywords: Action Recognition; Knowledge Distillation ; Action Recognition in the Dark

Nguyễn Tuấn Linh, Nhận học vị Tiến sỹ tại Học viện Công nghệ Bưu chính Viễn Thông năm 2022. Hiện đang công tác tại Trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên. Lĩnh vực nghiên cứu: Kỹ thuật máy tính, điện toán tỏa khắp, các mô hình học máy, chắt lọc tri thức và công nghệ cảm biến cho các ứng dụng chăm sóc sức khỏe.
Email: ntlinh.khmt@ictu.edu.vn

Hoàng Anh Tú, Tốt nghiệp đại học ngành Tin học, Đại học Khoa học Tự nhiên, Đạo học Quốc gia Hà Nội năm 2000. Nhận bằng Thạc Sỹ tại Đại học Aalborg năm 2003. Hiện tại đang là nghiên cứu sinh ngành Kỹ thuật máy tính tại Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Học sâu, chắt lọc tri thức, điện toán biên, mô hình sinh.
Email: hatu@mst.gov.vn