**Đếm lặp lại video bất khả tri lớp trong tự nhiên**

# Abstract

Một phương pháp để ước tính khoảng thời gian mà một hành động được lặp lại trong video. Điểm mấu chốt của phương pháp này nằm ở việc hạn chế mô-đun dự đoán chu kỳ sử dụng tính tương tự theo thời gian như một nút thắt biểu diễn trung gian cho phép sự tổng quát hóa lặp lại không nhìn thấy trong các video tự nhiên. Mô hình này, được gọi là RepNet, với tập dữ liệu tổng hợp được tạo từ bộ sưu tập video lớn không gắn nhãn bằng cách lấy mẫu các clip ngắn có độ dài khác nhau và lặp lại chúng với các khoảng thời gian và số lượng khác nhau. Sự kết hợp giữa dữ liệu tổng hợp và một mô hình mạnh mẽ nhưng bị hạn chế, cho phép dự đoán các định kỳ theo kiểu bất khả tri (chưa biết rõ). Mô hình về cơ bản vượt qua hiệu suất hiện đại trên các điểm chuẩn của thành phố thời kỳ hiện tại (PERTUBE) và đếm số lần lặp lại (QUVA). Một tập dữ liệu đầy thử thách mới có tên là Countix (lớn hơn ∼90 lần so với các tập dữ liệu hiện có). Trang web của dự án: <https://sites.google.com/view/repnet>.

# 1. Introduction

Hình dung những cảnh đời thường nhất - một người đang ăn trong quán cà phê. Họ có thể khuấy đường trong cà phê trong khi nhai thức ăn và gõ chân theo nhạc. Người này đang thực hiện song song ít nhất ba hoạt động định kỳ. Các hành động và quy trình lặp đi lặp lại rất phổ biến trong cuộc sống hàng ngày của chúng ta. Chúng bao gồm các chu kỳ hữu cơ, chẳng hạn như nhịp tim và nhịp thở, thông qua lập trình và sản xuất, đến các chu kỳ hành tinh như chu kỳ ngày-đêm và các mùa. Do đó, nhu cầu nhận dạng các lần lặp lại trong video là phổ biến và một hệ thống có thể xác định và đếm số lần lặp lại trong video sẽ mang lại lợi ích cho bất kỳ hệ thống tri giác nào nhằm mục đích quan sát và hiểu thế giới của chúng ta trong một khoảng thời gian dài.

Sự lặp lại cũng thú vị vì những lý do sau:

(1) Thường có một ý định hoặc một nguyên nhân thúc đẩy đằng sau một cái gì đó xảy ra nhiều lần;

(2) Cùng một sự kiện có thể được quan sát lại nhưng có sự khác biệt nhỏ;

(3) Có thể có những thay đổi dần dần trong cảnh do kết quả của những lần lặp lại này;

(4) Chúng cung cấp cho chúng ta các đơn vị hành động rõ ràng, một chuỗi con trong hành động có thể được phân đoạn kịp thời (ví dụ: nếu bạn đang cắt một củ hành, đơn vị hành động là hành động thao tác được lặp lại để tạo ra các lát bổ sung).

Do những lý do trên, bất kỳ tác nhân nào tương tác với thế giới sẽ được hưởng lợi rất nhiều từ một hệ thống như vậy. Hơn nữa, việc đếm số lần lặp lại là thích hợp cho nhiều ứng dụng thị giác máy tính; chẳng hạn như đếm số lần một bài tập thể dục đã được thực hiện, đo lường các sự kiện sinh học (như nhịp tim), v.v.

Tuy nhiên, nghiên cứu về video định kỳ còn hạn chế, có thể do thiếu tập dữ liệu lặp lại video được gắn nhãn quy mô lớn. Ngược lại, để nhận dạng hành động, có những bộ dữ liệu quy mô lớn, như Kinetics, nhưng bộ dữ liệu của chúng ở quy mô lớn được hữu dụng nhờ sự sẵn có của các từ khóa/văn bản được liên kết với video. Thật không may, hiếm khi video được gắn nhãn chú thích liên quan đến hoạt động lặp lại vì văn bản có nhiều khả năng mô tả nội dung ngữ nghĩa hơn. Vì lý do này, nên sử dụng tập dữ liệu với các nhãn hành động ngữ nghĩa thường được sử dụng để nhận dạng hành động (Kinetics) và chọn thủ công video về các lớp đó có chuyển động tuần hoàn (bóng nảy, vỗ tay, v.v.). Bước tiếp theo tiến hành gắn nhãn các video đã chọn với số lần lặp lại có trong mỗi clip.

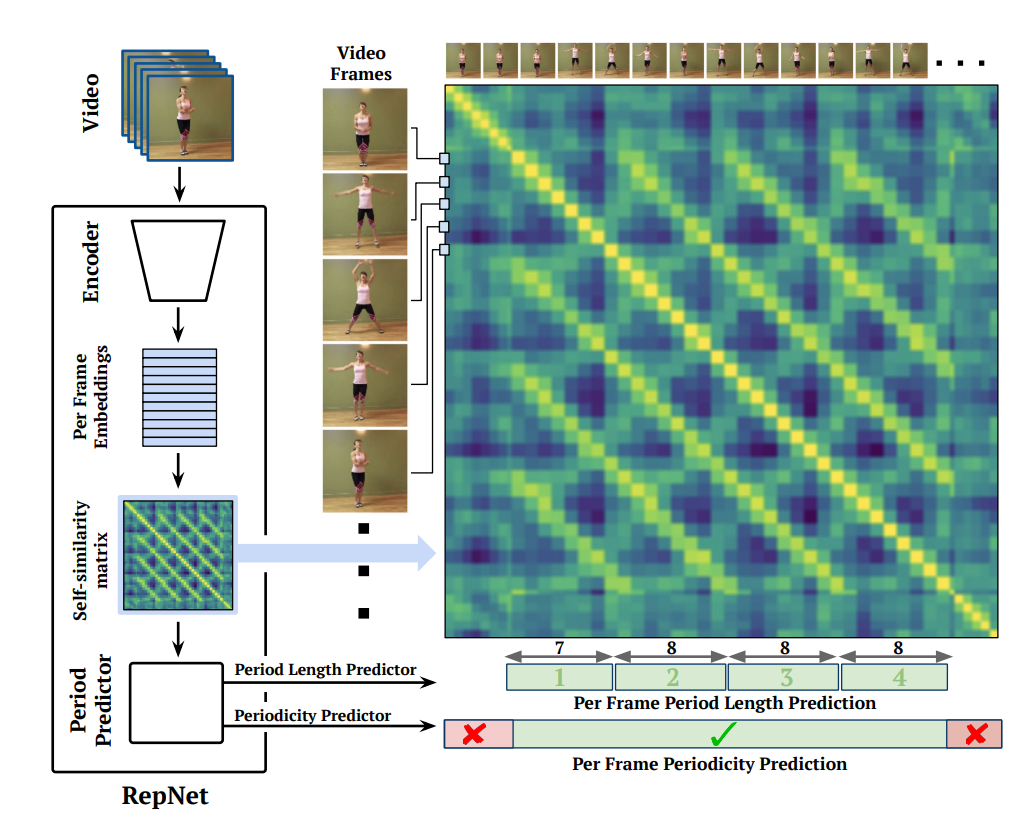
Việc gắn nhãn thủ công giới hạn số lượng video có thể được chú thích - việc gắn nhãn rất tẻ nhạt và tốn kém do tính chất chi tiết về thời gian của nhiệm vụ. Để giảm bớt lượng dữ liệu đào tạo, đề xuất một phương pháp tạo video lặp lại tổng hợp bằng cách lặp lại các clip từ các video hiện có với các khoảng thời gian khác nhau. Vì chúng ta đang tổng hợp định cỡ các video này nên cũng có chú thích chính xác cho khoảng thời gian và số lần lặp lại trong video, có thể được sử dụng cho các mô hình đào tạo sử dụng phương pháp học có giám sát. Như đã thấy trong công việc của mình, những video tổng hợp như vậy không thể nắm bắt được tất cả các sắc thái của các video lặp lại thực sự và dễ bị các mô hình học sâu dung lượng cao làm cho quá khớp. Để giải quyết vấn đề này, đề xuất chiến lược tăng dữ liệu cho video tổng hợp để các mô hình được đào tạo về chúng chuyển sang video thực có lặp lại. Chúng tôi sử dụng kết hợp dữ liệu thực và dữ liệu tổng hợp để phát triển mô hình của mình.

Trong bài viết này, mục tiêu của chúng ta là một mô hình duy nhất hoạt động cho nhiều lớp video định kỳ và trên thực tế, cho cả các lớp video không được nhìn thấy trong quá trình đào tạo. Chúng ta đạt được điều này bằng cách sử dụng một đại diện trung gian khuyến khích tạo thành các lớp không nhìn thấy. Biểu diễn này- ma trận tự tương tự về thời gian- được sử dụng để dự đoán khoảng thời gian mà một hành động được lặp lại trong video. Biểu diễn chung này được sử dụng trên các loại video lặp lại khác nhau cho phép tổng quát hóa mong muốn. Ví dụ: cho dù một người đang chống đẩy hay một đứa trẻ đang đu trong sân chơi, thì ma trận tự tương tự là phép tham số hóa được chia sẻ từ đó suy ra số lần lặp lại. Nút thắt cổ chai cực độ này (số lượng kênh trong bản đồ tính năng giảm từ 512 xuống còn 1) cũng hỗ trợ việc tổng hợp hóa từ dữ liệu tổng hợp thành dữ liệu thực. Ưu điểm khác của cách biểu diễn này là khả năng diễn giải của mô hình được đưa vào kiến ​​trúc mạng khi chúng ta buộc mạng chỉ dự đoán khoảng thời gian từ ma trận tự tương tự, trái ngược với việc suy ra khoảng thời gian từ các đặc trưng chiều cao tiềm ẩn.

Tập trung vào hai nhiệm vụ:

(i) Đếm số lần lặp lại và xác định số lần lặp lại trong video. Diễn đạt lại vấn đề này như ước tính độ dài trên mỗi chu kỳ khung hình, sau đó chuyển đổi chúng thành số lần lặp lại;

(ii) Phát hiện định kỳ, xác định xem khung hiện tại có phải là một phần của mô hình thời gian lặp lại hay không. Chúng tôi tiếp cận điều này như một vấn đề phân loại nhị phân trên mỗi khung. Giải thích trực quan về các nhiệm vụ này và tổng quan về giải pháp được hiển thị trong Hình 1.



Hình 1: Chúng tôi trình bày RepNet, sử dụng ma trận tự tương tự thời gian làm lớp trung gian để dự đoán độ dài chu kỳ và tính chu kỳ của mỗi khung hình trong video.

Những đóng góp chính của chúng tôi cho bài báo này là:

(i) RepNet, một kiến trúc mạng nơ-ron được thiết kế để đếm số lặp lại trong video một cách tự nhiên.

(ii) Một phương pháp để tạo và tăng cường các video lặp lại tổng hợp từ các video không được gắn nhãn.

(iii) Bằng cách đào tạo RepNet về tập dữ liệu tổng hợp, chúng tôi làm tốt hơn các phương pháp SOTA về cả các nhiệm vụ đếm lặp lại và phát hiện định kỳ so với các điểm chuẩn hiện có một biên độ đáng kể.

(iv) Tập dữ liệu đếm lặp lại video mới, Countix, lớn hơn ∼ 90 lần so với tập dữ liệu lớn nhất trước đó.

# 2. Related Work

Phép tương tự. Ý tưởng sử dụng hình ảnh địa phương và các điểm tương đồng về mặt không gian đã được khám phá trong [40] để đối sánh mẫu trong hình ảnh và video. Khớp với sự trừu tượng của các điểm tự tương đồng, thay vì trực tiếp các đặc điểm hình ảnh, đã kích hoạt tính tổng quát hóa. Chúng tôi xây dựng dựa trên cái nhìn sâu sắc này trong công việc của chúng tôi.

Ước tính định kỳ. Trích xuất tính tuần hoàn (phát hiện chuyển động tuần hoàn) và chu kỳ bằng cách tận dụng tự tương quan trong chuỗi thời gian là một vấn đề được nghiên cứu kỹ lưỡng [43, 48]. Ước tính chu kỳ trong video đã được thực hiện bằng cách sử dụng các biểu đồ xung quanh tương quan tự động [9] hoặc biến đổi Wavelet trên các tính năng được thiết kế thủ công bắt nguồn từ luồng quang học [37]. Chuyển động tuần hoàn được trích xuất đã hỗ trợ nhiều nhiệm vụ bao gồm tái tạo 3D [4, 29] và phân loại loài chim [28]. Tính chu kỳ đã được sử dụng cho các ứng dụng khác nhau [9, 32, 34, 38] bao gồm cả phân loại mẫu thời gian [35].

Ma trận tự tương tự theo thời gian (TSM). TSM là những đại diện hữu ích để ghi nhận hành động của con người [21, 24, 44] và phân tích dáng đi [5, 6] do chúng có khả năng chống lại những thay đổi lớn về góc nhìn khi được ghép nối với các đại diện tính năng thích hợp. TSM dựa trên hình ảnh qũy đạo dày đặc được cải tiến [49] được sử dụng trong [33] để xác định không giám sát các phân đoạn định kỳ trong video bằng cách sử dụng các bộ lọc đặc biệt. Không giống như những cách tiếp cận này, chúng tôi sử dụng TSM như một lớp trung gian trong kiến ​​trúc mạng nơ-ron end-to-end, hoạt động như một nút cổ chai thông tin. Đồng thời, [22] đã đề xuất một kiến ​​trúc phức hợp để phát hiện tính tuần hoàn trong video.

Dữ liệu đào tạo tổng hợp. Việc sử dụng dữ liệu đào tạo tổng hợp trong thị giác máy tính đang trở nên phổ biến hơn. Dán các bản vá đối tượng lên hình ảnh thực đã được chứng minh là có hiệu quả như dữ liệu đào tạo để phát hiện đối tượng [12, 15, 45] và ước tính tư thế con người [46]. Việc kết hợp nhiều video hoặc nhiều hình ảnh với nhau rất hữu ích cho việc tạo ra dữ liệu đào tạo tổng hợp cho các nhiệm vụ cụ thể [2] cũng như chính quy hóa các mô hình học sâu [53, 54]. Dữ liệu tổng hợp cho phép đếm lặp lại huấn luyện lần đầu tiên được đề xuất bởi [27]. Họ giới thiệu tập dữ liệu về các mẫu lặp lại tổng hợp và sử dụng tập dữ liệu này để đào tạo mô hình đếm dựa trên học sâu. Tuy nhiên, dữ liệu mà họ sử dụng để đào tạo bao gồm các mẫu ngẫu nhiên được thiết kế bằng tay không có vẻ thực tế. Như được trình bày trong [37], những mẫu này không đủ đa dạng để nắm bắt được tất cả các sắc thái của sự lặp lại trong video thực. Thay vào đó, chúng tôi đề xuất tạo tập dữ liệu đào tạo tổng hợp về các lần lặp lại video thực tế từ tập dữ liệu video hiện có.

Đếm trong Thị giác Máy tính. Đếm đối tượng và người trong hình ảnh [3, 7, 26, 30, 52] là một khu vực hoạt động trong thị giác máy tính. Mặt khác, việc đếm số lần lặp lại bằng video [27, 37] đã ít thu hút được sự quan tâm của cộng đồng trong thời đại học sâu. Chúng tôi xây dựng dựa trên ý tưởng [27] dự đoán chu kỳ (độ dài chu kỳ), mặc dù [27] không sử dụng TSM.

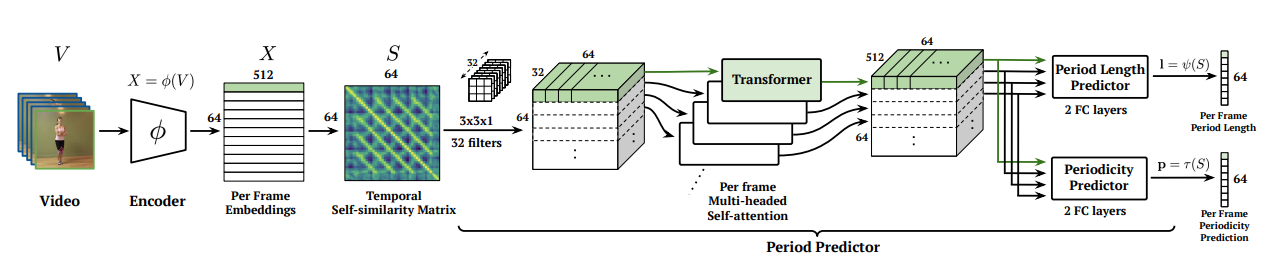
Nhiệm vụ chi tiết tạm thời. Đếm lặp lại và phát hiện định kỳ là các nhiệm vụ chi tiết về thời gian như xác định hành động tạm thời [8, 41], phân loại pha trên mỗi khung hình [11] và dự đoán trong tương lai [10]. Chúng tôi tận dụng các giao diện được sử dụng trước đây để thu thập tập dữ liệu bản địa hóa hành động, chẳng hạn như [16, 25, 42] để tạo Countix tập dữ liệu lặp lại của chúng tôi. Thay vì chú thích các phân đoạn ngữ nghĩa, chúng tôi gắn nhãn mức độ của các phân đoạn định kỳ trong video và số lần lặp lại trong mỗi phân đoạn.

# 3. RepNet Model

Trong phần này, giới thiệu kiến trúc RepNet, bao gồm hai thành phần đã học, bộ mã hóa và bộ dự đoán chu kỳ, với một lớp tự tương tự theo thời gian ở giữa chúng.

Giả sử chúng ta được cho một video V = [v1, v2, ..., vN] là một chuỗi N khung. Đầu tiên, chúng tôi cung cấp video V vào bộ mã hóa hình ảnh φ là X = φ(V) để tạo ra các nhúng trên mỗi khung hình X = [x1, x2, ..., xN]T. Sau đó, sử dụng phép nhúng X chúng ta thu được ma trận tự tương tự S bằng cách tính toán các điểm tương đồng theo cặp Sij giữa tất cả các cặp nhúng.

Cuối cùng, S được đưa vào mô-đun dự báo chu kỳ, mô-đun này xuất ra hai phần tử cho mỗi khung: ước tính độ dài chu kỳ Ɩ = ψ (S) và điểm tuần hoàn p = τ (S). Độ dài khoảng thời gian là tốc độ lặp lại diễn ra trong khi điểm số chu kỳ cho biết liệu khung hình có nằm trong một phần định kỳ của video hay không. Kiến trúc tổng thể có thể được xem trong Hình 1 và một phiên bản chi tiết hơn có thể được xem trong Hình 2.



Hình 2: Kiến trúc RepNet. Các tính năng được tạo ra bởi một khung video duy nhất được đánh dấu bằng màu xanh lá cây trên toàn mạng.

## 3.1. Encoder

Bộ mã hóa φ của chúng tôi bao gồm ba thành phần chính:

**Bộ giải nén tính năng tích chập** (Convolutional feature extractor): Chúng tôi sử dụng kiến ​​trúc ResNet-50 [19] làm mạng (CNN) cơ sở để trích xuất các tính năng phức hợp 2D từ các khung hình vi riêng lẻ của video đầu vào. Các khung này có kích thước 112 × 112 × 3. Chúng tôi sử dụng đầu ra của lớp conv4\_block3 để có một bản đồ tính năng địa lý 2D không gian lớn hơn. Các tính năng trên mỗi khung hình thu được có kích thước 7 × 7 × 1024.

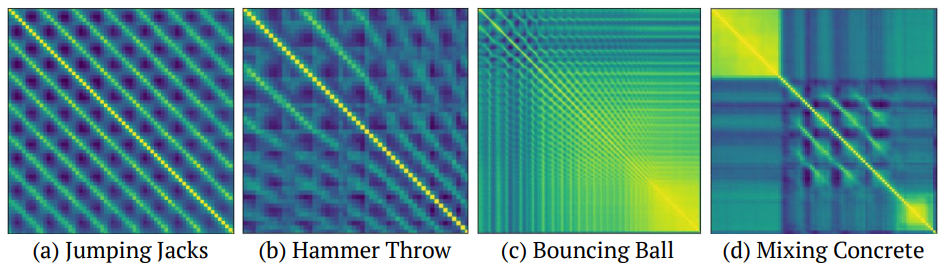
**Bối cảnh theo thời gian**: Chúng tôi chuyển các tính năng phức hợp này qua một lớp phức hợp 3D để thêm thông tin thời gian cục bộ vào các tính năng trên mỗi khung hình. Chúng tôi sử dụng 512 bộ lọc có kích thước 3 × 3 × 3 với kích hoạt ReLU với tỷ lệ giãn nở là 3. Bối cảnh thời gian giúp lập mô hình chuyển động ngắn hạn [13, 51] và cho phép mô hình phân biệt giữa các khung trông giống nhau nhưng khác chuyển động (ví dụ: tay di chuyển lên hoặc xuống trong khi tập thể dục).

**Giảm kích thước**: Chúng tôi giảm kích thước của các đặc điểm không gian-thời gian được trích xuất bằng cách sử dụng Global 2D Max-pooling trên các kích thước không gian và tạo ra các vectơ nhúng xi tương ứng với mỗi khung hình vi trong video

## 3.2. Ma trận tự tương tự theo thời gian (TSM)

Sau khi thu được các phép nhúng tiềm ẩn xi cho mỗi khung vi, chúng tôi xây dựng ma trận tự tương tự S bằng cách tính toán tất cả các điểm tương tự theo cặp Sij = f(xi, xj) giữa các cặp nhúng xi và xj, trong đó f(.) là hàm tương tự. Chúng tôi sử dụng phủ định của khoảng cách euclide bình phương làm hàm tương tự, f (a, b) = -|| a - b ||2, theo sau là phép toán softmax theo hàng.

Vì TSM chỉ có một kênh, nó hoạt động như một nút thắt cổ chai thông tin ở giữa mạng của chúng tôi và cung cấp sự chính quy hóa. TSM cũng làm cho mô hình có thể diễn giải được theo thời gian, điều này mang lại hiểu biết sâu hơn về các dự đoán được thực hiện bởi mô hình. Một số ví dụ có thể được xem trong Hình 3.



Hình 3: Sự đa dạng của ma trận tự tương tự theo thời gian được tìm thấy trong các video lặp lại trong thế giới thực (màu vàng có nghĩa là độ tương đồng cao, màu xanh lam có nghĩa là độ tương đồng thấp).

(a) Chuyển động tuần hoàn lặp lại đồng đều (nhảy dây)

(b) Lặp lại với gia tốc (vận động viên thực hiện ném búa)

(c) Lặp lại với chu kỳ giảm dần (một quả bóng nảy mất tốc độ do bị đập xuống đất nhiều lần)

(d) Chuyển động lặp lại trước đó và không chuyển động tiếp theo (chờ trộn bê tông, trộn bê tông, dừng trộn).

Một mô hình phức tạp là cần thiết để dự đoán chu kỳ và tính tuần hoàn từ các ma trận tự tương tự đa dạng như vậy.

## 3.3. Dự đoán chu kỳ

Mô-đun cuối cùng của RepNet là công cụ dự đoán thời gian. Mô-đun này chấp nhận ma trận tự tương tự S = [s1, s2, ..., sN]T trong đó mỗi hàng si là biểu diễn tự tương tự trên mỗi khung và tạo ra hai đầu ra: ước tính độ dài chu kỳ trên mỗi khung hình Ɩ = ψ( S), và phân loại tuần hoàn nhị phân trên mỗi khung p = τ(S). Lưu ý rằng cả Ɩ và p đều là vectơ và các phần tử của chúng là các dự đoán trên mỗi khung (tức là Ɩi là độ dài chu kỳ dự đoán cho khung thứ i).

Kiến trúc của mô-đun dự báo chu kỳ có thể được xem trong Hình 2. Lưu ý rằng các mô-đun dự báo ψ và τ có chung một kiến ​​trúc và trọng số cho đến giai đoạn phân loại cuối cùng. Quy trình xử lý được chia sẻ bắt đầu với 32 bộ lọc tích chập 2D có kích thước 3 × 3, tiếp theo là lớp transformer[47] sử dụng sự chú ý từ nhiều phía (multi-headed attention) với các phép nhúng vị trí có thể huấn luyện dưới dạng một biến độ dài 64 được học qua đào tạo. Chúng tôi sử dụng 4 đầu với 512 chiều trong transformer với mỗi đầu có kích thước 128 chiều. Sau đường ống được chia sẻ, chúng ta có hai bộ phân loại, bộ phân loại độ dài chu kỳ ψ và bộ phân loại tuần hoàn τ. Mỗi lớp bao gồm hai lớp được kết nối đầy đủ có kích thước 512.

## 3.4. Losses

Bộ phân loại tuần hoàn của chúng tôi τ xuất ra mỗi khung phân loại định kỳ pi và sử dụng mất mát phân loại nhị phân (binary cross-entropy) để tối ưu hóa. Công cụ ước tính độ dài chu kỳ của chúng tôi ψ đầu ra cho mỗi ước tính độ dài chu kỳ khung Ɩ i ∈ L trong đó các lớp là độ dài chu kỳ rời rạc L = {2, 3, ..., N/2} trong đó N là số khung đầu vào. Chúng tôi sử dụng mục tiêu phân loại nhiều lớp (softmax cross-entropy) để tối ưu hóa mô hình của mình. Đối với tất cả các thử nghiệm của chúng tôi, chúng tôi sử dụng N = 64. Chúng tôi lấy mẫu video đầu vào với các tốc độ khung hình khác nhau như mô tả bên dưới để dự đoán độ dài khoảng thời gian lớn hơn.

## 3.5. Sự suy luận

Việc suy ra số lần lặp lại mạnh mẽ cho một video nhất định yêu cầu hai thao tác chính:

Đếm từ bước dự đoán độ dài chu kỳ: Chúng tôi lấy mẫu các cửa sổ không chồng chéo liên tiếp của N khung và cung cấp nó làm đầu vào cho RepNet, xuất ra số pi theo chu kỳ trên mỗi khung và độ dài chu kỳ Ɩi. Chúng tôi xác định số lượng mỗi khung hình làpi/ Ɩi . Tổng Số lần lặp lại được tính bằng tổng của tất cả các lần đếm trên mỗi khung hình: . Bộ dữ liệu đánh giá để đếm số lần lặp lại chỉ có các phân đoạn định kỳ. Do đó, chúng tôi đặt pi thành 1 làm giá trị mặc định cho các thử nghiệm đếm.

Đánh giá đa tốc độ: Vì mô hình của chúng tôi có thể dự đoán độ dài khoảng thời gian lên đến 32, để bao gồm độ dài khoảng thời gian dài hơn nhiều, chúng tôi lấy mẫu video đầu vào với các tốc độ khung hình khác nhau. (tức là chúng tôi phát video ở tốc độ 1 ×, 2 ×, 3 × và 4 ×). Chúng tôi chọn tốc độ khung hình có điểm cao nhất trong khoảng thời gian dự đoán. Điều này tương tự như những gì [27] thực hiện tại thời điểm thử nghiệm.

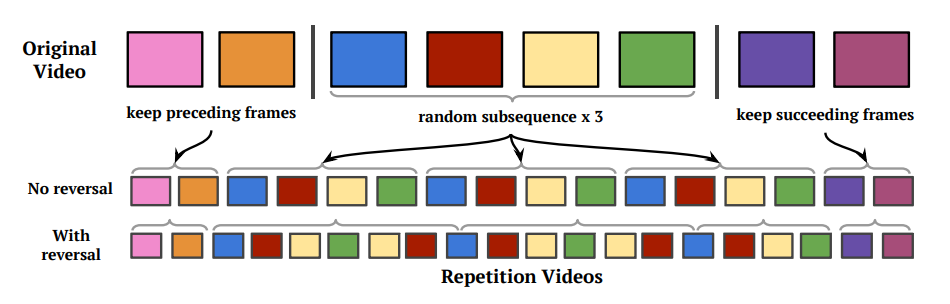
# 4. Đào tạo với các lần lặp lại tổng hợp

Một phương pháp tiếp cận có giám sát tiềm năng để ước tính khoảng thời gian sẽ là thu thập một tập hợp lớn các video đào tạo định kỳ và chú thích phần đầu và phần cuối của mỗi kỳ trong tất cả các hành động lặp lại. Tuy nhiên, việc thu thập một tập dữ liệu như vậy rất tốn kém do tính chất chi tiết của nhiệm vụ.

## Là một giải pháp thay thế rẻ hơn và có thể mở rộng hơn, chúng tôi đề xuất một chiến lược đào tạo sử dụng các lần lặp lại được tạo tổng hợp bằng cách sử dụng các video không được gắn nhãn trong tự nhiên (ví dụ: YouTube). Chúng tôi tạo các video định kỳ tổng hợp bằng cách sử dụng các video được chọn ngẫu nhiên và dự đoán độ dài chu kỳ và chu kỳ trên mỗi khung hình. Tiếp theo, chúng tôi sẽ giải thích cách chúng tôi tạo ra số lần lặp lại tổng hợp và giới thiệu tính năng nâng cao chuyển động của máy ảnh rất quan trọng để đào tạo các mô hình đếm hiệu quả từ video tổng hợp.

## 4.1. Tổng hợp các video lặp lại

Với một tập hợp lớn các video không được gắn nhãn, chúng tôi đề xuất một cách tiếp cận đơn giản nhưng hiệu quả để tạo các video lặp lại tổng hợp (thể hiện trong Hình 4) từ chúng. Lợi thế của việc sử dụng video thực để tạo dữ liệu tổng hợp là dữ liệu đào tạo gần với video lặp lại thực hơn nhiều so với việc sử dụng các mẫu tổng hợp. Một lợi thế khác của việc sử dụng video thực là việc sử dụng một bộ dữ liệu lớn như Kinetics đảm bảo rằng tính đa dạng của dữ liệu mà mô hình nhìn thấy là rất lớn. Điều này cho phép chúng tôi đào tạo các mô hình phức tạp lớn có thể hoạt động trên các video lặp lại thực tế.



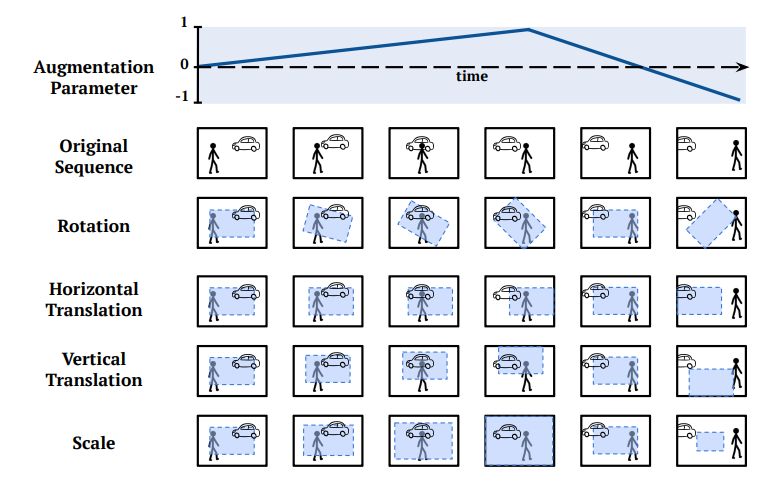
Hình 4: Quy trình tạo dữ liệu tổng hợp của chúng tôi tạo ra các video có sự lặp lại từ bất kỳ video nào. Chúng tôi lấy mẫu ngẫu nhiên một phần của video mà chúng tôi lặp lại N lần để tạo ra các video lặp lại tổng hợp. Thêm chi tiết trong Phần 4

Quy trình của chúng tôi bắt đầu bằng việc lấy mẫu một video V ngẫu nhiên từ tập dữ liệu các video. Chúng tôi sử dụng tập huấn luyện của Kinetics [23] mà không có bất kỳ nhãn nào. Sau đó, chúng tôi lấy mẫu một clip C có độ dài ngẫu nhiên P khung hình từ V. Đoạn clip C này được lặp lại K lần (trong đó K> 1) để mô phỏng các video có số lần lặp lại. Chúng tôi ghép ngẫu nhiên clip đã đảo ngược trước khi lặp lại để mô phỏng các hành động trong đó chuyển động được thực hiện ngược lại trong khoảng thời gian (như nhảy dây). Sau đó, chúng tôi sắp xếp trước và nối các khung lặp lại với các đoạn không lặp lại khác từ V, các đoạn này ngay trước và sau C, tương ứng. Độ dài của các đoạn không theo chu kỳ này được chọn ngẫu nhiên và cũng có thể bằng 0.

Thao tác này đảm bảo rằng có cả phân đoạn định kỳ và không định kỳ trong video được tạo. Cuối cùng, mỗi khung hình trong phần lặp lại của video đã tạo được gán nhãn độ dài chu kỳ P. Một nhãn định kỳ cũng được tạo cho biết khung hình nằm bên trong hay bên ngoài phần lặp lại của video đã tạo.

## 4.2. Tăng cường chuyển động máy ảnh

Một bước quan trọng trong quá trình tạo video tổng hợp là tăng cường chuyển động của camera (thể hiện trong Hình 5). Mặc dù không khả thi khi dự đoán chế độ xem của một máy ảnh chuyển động tùy ý mà không biết cấu trúc 3D, các bộ phận bị che khuất và nguồn ánh sáng trong cảnh, chúng tôi có thể ước tính nó bằng cách sử dụng các phép biến đổi hình ảnh affine. Ở đây chúng tôi xem xét chuyển động liên kết của khung xem trên video, bao gồm các thay đổi tạm thời về xoay, dịch và tỷ lệ. Như chúng tôi sẽ trình bày trong phần 6, khi chúng tôi đào tạo mà không có các phần bổ sung này, tổn thất đào tạo nhanh chóng giảm xuống nhưng mô hình không chuyển sang video lặp lại thực sự. Theo kinh nghiệm, chúng tôi nhận thấy việc nâng cao chuyển động của máy ảnh là một phần quan trọng trong việc đào tạo các mô hình hiệu quả bằng video tổng hợp.



Hình 5: Chuyển động của camera. Chúng tôi thay đổi các thông số nâng cao cho từng loại máy ảnh di chuyển trơn tru theo thời gian thay vì lấy mẫu ngẫu nhiên chúng một cách độc lập cho từng khung hình. Điều này đảm bảo rằng trình tự tăng cường vẫn giữ được sự mạch lạc theo thời gian hiện diện tự nhiên trong video.

# Để tăng cường chuyển động của máy ảnh, chúng tôi tạm thời thay đổi các thông số cho các loại chuyển động khác nhau một cách liên tục khi video tiếp tục. Ví dụ, chúng tôi thay đổi góc quay trơn tru theo thời gian. Điều này đảm bảo rằng video nhất quán về mặt thời gian ngay cả sau khi nâng cấp. Hình 5 minh họa cách thông số tăng thời gian điều khiển khung xem (được hiển thị trong hình chữ nhật màu xanh lam) cho từng loại chuyển động. Điều này dẫn đến các video có ít bản sao gần như trùng lặp trên các phân đoạn lặp lại.

# 5. Countix Dataset

Các bộ dữ liệu hiện có để đếm số lần lặp lại video [27, 37] hầu hết được sử dụng cho mục đích thử nghiệm, chủ yếu là do kích thước hạn chế của chúng. Tiêu chuẩn thử thách và gần đây nhất cho nhiệm vụ này là tập dữ liệu lặp lại QUVA [37] bao gồm các video lặp lại thực tế với hiện tượng khớp, chuyển động của camera và thay đổi về tốc độ của các hành động lặp lại. Nó bao gồm 100 video kiểm tra bất khả tri theo lớp, được chú thích với số lượng các hành động lặp lại. Mặc dù đầy thách thức, nhưng kích thước hạn chế của nó làm cho nó khó có thể bao gồm các loại lặp lại ngữ nghĩa đa dạng. Ngoài ra, đào tạo mô hình sâu có giám sát với quy mô dữ liệu này là không khả thi.

Để tăng tính đa dạng về ngữ nghĩa và mở rộng quy mô của bộ dữ liệu đếm, chúng tôi giới thiệu bộ dữ liệu Countix: bộ dữ liệu trong thế giới thực gồm các video lặp lại được thu thập trong môi trường hoang dã (tức là YouTube) bao gồm một loạt các cài đặt ngữ nghĩa với những thách thức đáng kể như máy ảnh và chuyển động của đối tượng, tập hợp các khoảng thời gian và số đếm đa dạng, và những thay đổi về tốc độ của các hành động lặp lại.

Countix bao gồm các video lặp đi lặp lại về các hoạt động tập luyện (ngồi xổm, kéo xà, luyện dây chiến đấu, tập thể dục cánh tay), các động tác nhảy (đánh đàn, nắm đấm), chơi nhạc cụ (chơi ukulele), sử dụng các dụng cụ nhiều lần (búa đập vào đồ vật, cưa máy cắt gỗ, cắt hành tây), biểu diễn nghệ thuật (nhảy hula hooping, tung hứng bóng đá), thể thao (chơi bóng bàn và quần vợt) và nhiều môn khác. Hình 6 minh họa một số ví dụ từ tập dữ liệu cũng như sự phân bố số lần lặp lại và độ dài chu kỳ. Tập hợp tập dữ liệu: Tập dữ liệu Countix là một tập con của tập dữ liệu Kinetics [23] được chú thích bằng các phân đoạn của các hành động được duyệt lại và số lượng tương ứng. Trong quá trình thu thập, trước tiên, chúng tôi chọn một cách thủ công một tập hợp con các lớp từ Kinetics có khả năng lặp lại cao hơn trong chúng, ví dụ: nhảy dây, thái hành tây, v.v., chứ không phải là các lớp học như đấu vật đối đầu hoặc đấu vật cá sấu.

Chúng tôi cộng gộp các nhãn cho các phân đoạn lặp lại và số lượng cho các lớp đã chọn. Giao diện được sử dụng tương tự như giao diện thường được sử dụng để đánh dấu các phân đoạn thời gian để nhận dạng hành động chi tiết [16, 36]. Trước tiên, người chú thích được yêu cầu phân đoạn phần video chứa các đoạn lặp lại hợp lệ với số lượng rõ ràng. Các ký hiệu viên sau đó tiến hành đếm số lần lặp lại trong mỗi phân đoạn. Số lượng này đóng vai trò là nhãn cho toàn bộ clip. Chúng tôi từ chối các phân đoạn có sự trùng lặp không đáng kể trong phạm vi thời gian được đánh dấu bởi 3 chú thích khác nhau. Đối với các phân đoạn còn lại, chúng tôi sử dụng giá trị trung bình của số lượng chú thích và phạm vi phân đoạn làm chân lý cơ bản. Tập dữ liệu Countix lớn hơn khoảng 90 lần so với tập dữ liệu đếm số lần lặp lại lớn nhất trước đó (Bộ dữ liệu lặp lại QUVA). Các số liệu thống kê chi tiết có thể được xem trong Bảng 1. Bộ dữ liệu có sẵn trên trang web của dự án.

Lưu ý rằng chúng tôi giữ lại phần tách train / val / test từ tập dữ liệu Kinetics. Do đó, các mô hình được đào tạo trước với Kinetics có thể được sử dụng để đào tạo các mô hình đếm mà không bị rò rỉ dữ liệu.

# 6. Experiments

Chúng tôi bắt đầu bằng cách giải thích các điểm chuẩn hiện có và các chỉ số đánh giá được sử dụng trong việc đếm số lần lặp lại. Tiếp theo, chúng tôi trình bày một loạt các nghiên cứu cắt bỏ chứng minh các thành phần và lựa chọn thiết kế nào là quan trọng. Sau đó, chúng tôi đánh giá hiệu suất của mình trên các điểm chuẩn hiện có và cho thấy rằng RepNet rõ ràng làm tốt hơn các phương pháp SOTA về tính lặp lại và phát hiện định kỳ. Cuối cùng, thông qua phân tích định tính, chúng tôi mang lại cái nhìn sâu sắc hơn về mô hình của mình.

## 6.1. Điểm chuẩn và chỉ số đánh giá

Ở đây chúng tôi thảo luận về hai bộ dữ liệu điểm chuẩn được thiết lập để phát hiện định kỳ và đếm số lần lặp lại cùng với các số liệu đánh giá thường được sử dụng.

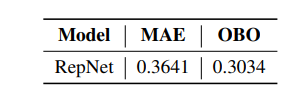
Phát hiện định kỳ: Tập dữ liệu chuẩn cho tác vụ này là tập dữ liệu PERTUBE [33], có các nhãn trên mỗi khung xác định tính chu kỳ, nếu khung là một phần của hành động lặp lại hay không. [33] xử lý vấn đề dưới dạng một nhiệm vụ phân loại nhị phân trên mỗi khung hình và báo cáo độ chính xác, thu hồi, điểm F1 và chồng chéo. Chúng tôi tuân theo các số liệu tương tự để đánh giá.

Đếm số lần lặp lại: Như đã thảo luận trong Phần 5, bộ dữ liệu QUVA [37] là bộ dữ liệu lớn nhất hiện có để đếm số lần lặp lại. Tài liệu hiện có sử dụng hai chỉ số chính để đánh giá số lần lặp lại trong video:

Lỗi đếm từng lỗi một (OBO). Nếu số lượng dự đoán nằm trong một số đếm của giá trị sự thật cơ bản, thì video đó được coi là được phân loại chính xác, nếu không, đó là phân loại sai. Lỗi OBO là tỷ lệ phân loại sai trên toàn bộ tập dữ liệu.

Sai số tuyệt đối trung bình (MAE) của số lượng. Chỉ số này đo lường sự khác biệt tuyệt đối giữa số lượng sự thật cơ bản và số lượng dự đoán, sau đó chuẩn hóa nó bằng cách chia cho số lượng sự thật cơ bản. Lỗi MAE được báo cáo là giá trị trung bình của sự khác biệt tuyệt đối được chuẩn hóa trên toàn bộ tập dữ liệu.

Cả trong các thử nghiệm cắt bỏ và so sánh hiện đại, chúng tôi tuân theo [27, 37] và báo cáo lỗi OBO và MAE qua bộ xác thực QUVA và Countix. Chúng tôi cũng cung cấp điểm số cuối cùng của bài kiểm tra Countix trong Bảng 7.



Bảng 7: Kết quả đếm trên bộ thử nghiệm Countix

## 6.2. Chi tiết triển khai

Chúng tôi triển khai phương pháp của mình trong Tensorflow [1]. Chúng tôi khởi tạo bộ mã hóa với trọng số từ một điểm kiểm tra ResNet-50 được đào tạo trước của ImageNet. Chúng tôi đào tạo mô hình cho 400 nghìn bước với tỷ lệ học tập là 6 × 10−6 với trình tối ưu hóa ADAM và kích thước hàng loạt của 5 video (mỗi video có 64 khung hình). Đối với tất cả các nghiên cứu cắt bỏ, chúng tôi đào tạo mô hình dựa trên dữ liệu thời gian lặp lại tổng hợp trừ khi có quy định khác. Thông tin chi tiết bổ sung được cung cấp trên trang web của dự án.

## 6.3. Hủy bỏ

Chúng tôi thực hiện một số biện pháp để biện minh cho các quyết định được đưa ra trong khi thiết kế RepNet.

Ma trận tự tương tự theo thời gian (TSM): Trong Bảng 2, chúng tôi so sánh tác động của việc thêm TSM vào mô hình. Model không có TSM áp dụng máy biến áp trực tiếp trên các nhúng trên mỗi khung hình do bộ mã hóa tạo ra. Ma trận tự tương tự theo thời gian cải thiện đáng kể hiệu suất trên tất cả các chỉ số và bộ dữ liệu xác thực cho dù chúng tôi đào tạo mô hình bằng cách sử dụng video lặp lại tổng hợp, video Coun tix thực hay kết hợp cả hai. Hơn nữa, lớp TSM giúp tổng quát hóa các video lặp lại thực sự ngay cả khi mô hình chỉ xem các video lặp lại tổng hợp (hàng 1 và 2 trong Bảng 2).

Nguồn dữ liệu đào tạo: Chúng tôi thay đổi các nguồn dữ liệu đào tạo trong Bảng 2 trong khi so sánh các video lặp lại tổng hợp của chúng tôi với các video thực từ tập dữ liệu Countix. Chúng tôi nhận thấy rằng Rep Net đạt được hiệu suất tương tự trên tập dữ liệu Countix khi được đào tạo với các video tổng hợp hoặc với các video lặp lại thực của tập dữ liệu Countix. Nhưng mô hình được đào tạo trên tập dữ liệu Countix kém hơn trên tập dữ liệu QUVA so với mô hình đào tạo về video lặp lại tổng hợp. Điều này cho thấy việc sử dụng tập dữ liệu lặp lại tổng hợp dẫn đến một mô hình cũng hoạt động cạnh tranh trên các lớp không nhìn thấy. Hiệu suất tốt nhất về lỗi OBO đạt được khi mô hình được đào tạo với cả hai bộ dữ liệu.

Kiến trúc dự báo chu kỳ thay thế: Trong Bảng 3, chúng tôi so sánh kiến ​​trúc Transformer với các mô hình trình tự hiện đại khác như LSTM và CNN tạm thời. Chúng tôi cũng so sánh nó với một mô hình sử dụng CNN 2D trên chính ma trận tự tương tự. Chúng tôi nhận thấy rằng kiến ​​trúc máy biến áp hoạt động tốt hơn các lựa chọn thay thế này.

Tăng cường chuyển động của máy ảnh: Trong Bảng 4, chúng tôi trình bày giá trị của việc tăng chuyển động của máy ảnh khi sử dụng tập dữ liệu lặp lại tổng hợp. Chúng tôi quan sát thấy rằng hiệu suất trên cả hai tập dữ liệu được cải thiện khi tỷ lệ mẫu trong lô có tăng cường chuyển động camera được tăng lên.

## 6.4. Đánh giá về điểm chuẩn

Chúng tôi so sánh hệ thống của mình với các phương pháp hiện đại nhất về phát hiện định kỳ và đếm số lần lặp lại trên các điểm chuẩn đã thiết lập được mô tả trong Phần 6.1. Phát hiện định kỳ. Chúng tôi báo cáo hiệu suất để đo phân loại chu kỳ bằng cách chọn ngưỡng tối đa hóa điểm F1. Như đã thực hiện trong [33], chúng tôi tính toán các chỉ số trên cơ sở mỗi video và tính điểm trung bình. Chúng tôi cũng báo cáo Diện tích dưới đường cong (AUC) của đường cong thu hồi độ chính xác độc lập với ngưỡng đã chọn

Mô hình của chúng tôi tạo ra AUC là 0,969. Chúng tôi làm tốt hơn công việc trước đó mà không sử dụng bất kỳ phương pháp tạo bộ lọc được thiết kế thủ công nào được đề cập trong [33] (xem Bảng 5). Mô hình của chúng tôi được đào tạo hoàn toàn dựa trên dữ liệu tổng hợp hoạt động hiệu quả cho nhiệm vụ phát hiện tính định kỳ trong video thực. Đếm số lần lặp lại. Trong Bảng 6, chúng tôi so sánh mô hình RepNet của chúng tôi với các mô hình trước đó và cho thấy nó hoạt động tốt hơn các phương pháp hiện có một biên độ đáng kể và do đó thiết lập một trạng thái mới nhất cho tập dữ liệu này. Kết quả thử nghiệm trên tập thử nghiệm của tập dữ liệu Countix chỉ ra rằng Rep Net là đường cơ sở hiệu quả cho nhiệm vụ đếm số lần lặp lại video (xem Bảng 7).

## 6.5. Phân tích định tính

Ma trận tự tương tự theo thời gian. TSM cung cấp cho chúng tôi những diễn giải có ý nghĩa về các dự đoán của mô hình. Nó cũng chứa thông tin bổ sung liên quan đến việc tăng và giảm tốc độ của hành động. Chúng tôi đưa ra một số ví dụ về ma trận tự tương tự trong Hình 3.

Nhúng 1D PCA. Chúng tôi cũng điều tra các loại chăn ga gối đệm đã học được sử dụng để sản xuất TSM. Trong Hình 7, chúng tôi chiếu vectơ 512 chiều lên 1 chiều bằng cách sử dụng thành phần chính đầu tiên của các lần nhúng trên mỗi khung hình cho mỗi video. Điều này cho thấy các mô hình gần như hình sin thú vị được tìm ra bằng cách nhúng theo thời gian. Chúng tôi vẽ biểu đồ các khung khi các nhúng đang thay đổi hướng và quan sát rằng các khung được truy xuất hiển thị người hoặc đối tượng ở trạng thái tương tự nhưng trong các khoảng thời gian khác nhau.

Lỗi đếm kép. Chúng tôi quan sát thấy rằng một chế độ thất bại phổ biến của mô hình của chúng tôi là đối với một số hành động (ví dụ: tung hứng quả bóng đá), nó dự đoán một nửa số lượng được báo cáo bởi một ký giả. Điều này xảy ra khi mô hình xem xét chuyển động của chân trái và chân phải để đếm trong khi mọi người có xu hướng xem xét chuyển động lên / xuống của quả bóng, dẫn đến việc mọi người đếm hai lần số lần lặp lại. Chúng tôi tin rằng những lỗi như vậy rất khó để cô lập theo cách bất khả tri của lớp. Nhưng chúng có thể được sửa dễ dàng bằng dữ liệu được gắn nhãn hoặc các phương pháp xử lý sau nếu ứng dụng được biết

# 7. Ứng dụng

Dự đoán thay đổi tốc độ của các lần lặp lại. Phương pháp này đưa vào một video clip và dự đoán khoảng thời gian của bất kỳ hành động lặp lại nào. Sự khác biệt liên tiếp của tỷ lệ dự đoán mã hóa tốc độ thay đổi tốc độ của các lần lặp lại. Theo dõi các thay đổi về tốc độ rất hữu ích cho các ứng dụng theo dõi tập thể dục, trong đó điều quan trọng là phải biết ai đó đang tăng tốc hay chạy chậm lại (Cột 4 trong Hình 8).

Ước tính tần suất của các quy trình từ video. Mô hình của chúng tôi có thể được sử dụng để dự đoán số lượng và tần suất của các hiện tượng lặp lại từ video, ví dụ: quá trình sinh học (nhịp tim). [50] đã trình bày một phương pháp để tiết lộ những thay đổi tinh vi bằng cách phóng đại sự khác biệt trong khung hình. Chúng tôi thấy rằng đầu ra từ hệ thống trên có thể được đưa trực tiếp vào mô hình của chúng tôi để dự đoán tần suất của những thay đổi này. Công cụ ước tính khoảng thời gian bất khả tri của lớp loại bỏ nhu cầu đào tạo rõ ràng về các video này. Trên trang web dự án của chúng tôi, chúng tôi hiển thị các ví dụ về việc đếm số lần lặp lại trên các video echo-cardiogram trông rất khác so với các video Kinetics.

Truy xuất chi tiết xuyên thời kỳ. Các cách nhúng đã học rất hữu ích để thực hiện truy xuất giữa các kỳ. Nói cách khác, các tính năng nắm bắt những điểm tương đồng hiện diện trong các thời kỳ khác nhau trong khi vẫn mã hóa sự khác biệt tinh tế giữa các khung hình trông giống nhau. Ví dụ về các truy xuất này được thể hiện trong Hình 7 và cột cuối cùng trong Hình 8.

Các lần lặp lại với thời gian dài hơn. Nhiều hiện tượng lặp lại xảy ra trên quy mô thời gian dài hơn (theo thứ tự ngày hoặc năm). Mặc dù mô hình của chúng tôi đã được đào tạo về video ngắn (∼10s), mô hình này vẫn có thể hoạt động trên video có các sự kiện định kỳ chậm bằng cách tự động chọn bước khung hình cao hơn. Trên trang web của dự án, chúng tôi hiển thị các video trong đó RepNet dự đoán khoảng thời gian của một ngày từ các video về trái đất được vệ tinh thu lại.

Hỗ trợ học tập đại diện video tự giám sát. Phương pháp học tập tự giám sát để nhúng video, ví dụ: Các mạng Shuffle and Learn [31], Odd-One-Out [14], DPC [17], TCC [11] và TCN [39] không được thiết kế để xử lý các lần lặp lại theo trình tự. RepNet có thể xác định các phần lặp lại và có thể giúp đào tạo về các video có phần lặp lại mà không cần sửa đổi các mục tiêu đã đề xuất.

# 8. Conclusion

Chúng tôi đã chỉ ra một sự kết hợp đơn giản giữa dữ liệu huấn luyện tổng hợp, cùng với kiến trúc sử dụng tính tương tự theo thời gian, dẫn đến một mô hình đếm lặp lại bất khả tri theo lớp mạnh mẽ. Mô hình này phát hiện thành công tính chu kỳ và dự đoán số lượng trên một loạt các tác nhân (vật thể, con người, động vật, trái đất) và cảm biến (máy ảnh tiêu chuẩn, siêu âm, kính hiển vi laze) và đã được đánh giá trên một bộ sưu tập lớn các video. Với điều này, chúng tôi đã giải quyết trường hợp lặp lại đơn giản và bước tiếp theo là xem xét các trường hợp phức tạp hơn như nhiều tín hiệu lặp lại đồng thời và sắp xếp thời gian của các phần lặp lại như trong các bước nhảy và âm nhạc.

Lời cảm ơn: Chúng tôi cảm ơn Aishwarya Gomatam, Anelia Angelova, Meghana Thotakuri, Relja Arandjelovic, Shefali Umrania, Sourish Chaudhuri và Vincent Van houcke đã giúp đỡ họ cho dự án này.

# (Appendix) Ruột thừa

Trên trang web dự án của chúng tôi, chúng tôi cung cấp hình ảnh trực quan về kết quả định tính, mẫu tập dữ liệu và Hình ảnh hóa PCA 1D.

# A. Kết quả Định tính

Tất cả các ví dụ dưới đây đã được tạo bằng một mô hình duy nhất được đào tạo chỉ với dữ liệu tổng hợp

# A.1. Đếm video có các cảm biến khác nhau

Chúng tôi cung cấp các ví dụ về mô hình của chúng tôi trên các cảm biến khác nhau:

Siêu âm tim. RepNet có thể ước tính nhịp tim từ các video siêu âm tim. Chúng tôi nhận thấy nhịp tim dự đoán gần với nhịp tim thực do chính thiết bị đo (liên kết đến video). Lưu ý cách hoạt động của cùng một kiểu máy trên các máy điện tâm đồ khác nhau.

Kính hiển vi laze. Chúng tôi đã tìm thấy video về các hiện tượng cali sinh học lặp lại ở [18], nơi họ quan sát các hiện tượng tế bào dưới kính hiển vi la-de, kết quả là tạo ra các mô hình xoắn ốc trong video. Chúng tôi nhận thấy rằng mô hình của chúng tôi hoạt động hiệu quả khi đo tốc độ mà các đường xoắn ốc đang quay. Mô hình cũng ghi lại sự thay đổi tốc độ trong quá trình được đo (liên kết đến video).

Video phóng đại Eulerian Mô hình của chúng tôi hoạt động dựa trên các video được tạo ra bằng cách sử dụng tính năng phóng đại Eulerian để làm nổi bật những thay đổi tinh tế theo thời gian [50]. Chúng tôi cho thấy RepNet có thể tin tưởng vào những video đó mà không cần đào tạo thêm (liên kết đến video).

# A.2. Thí nghiệm Vật lý với RepNet

Chúng tôi đưa ra các ví dụ về 2 video trong đó lắc các con lắc có độ dài khác nhau. Tỷ lệ khoảng thời gian có thể được sử dụng để dự đoán tỷ lệ chiều dài của các con lắc. Mô hình của chúng tôi có thể thay thế bước mà những người tiến hành thử nghiệm đo khoảng thời gian bằng đồng hồ bấm giờ. Chúng tôi tự mình tiến hành thí nghiệm từ video và tìm tỷ lệ khoảng thời gian của con lắc dài và ngắn sử dụng độ dài chu kỳ mà mô hình của chúng tôi dự đoán là 1,566. Dựa trên các phương trình vật lý về dao động của con lắc, tỷ lệ gần đúng dự kiến của các chu kỳ sử dụng giá trị xấp xỉ chiều dài của con lắc (từ pixel trong video) là 1,612. Chúng tôi cung cấp thông tin chi tiết trong các thử nghiệm (tại đây).

# A.3. Dự đoán khoảng thời gian có nhiều chế độ xem nhất quán

Chúng tôi kiểm tra mô hình của chúng tôi với các góc nhìn khác nhau ghi lại sự sụp đổ của Cầu Tacoma vào năm 1940 do cộng hưởng. Mô hình của chúng tôi phục hồi tần suất lặp lại từ các góc nhìn khác nhau một cách mạnh mẽ (liên kết đến video).

# A.4. Kiểm tra các thay đổi theo chu kỳ

RepNet lấy đầu vào là đại diện hình ảnh vệ tinh (do NASA Goddard phát hành) về lớp băng phủ trên Bắc Cực trong khoảng thời gian 25 năm và dự đoán khoảng thời gian này là khoảng 1 năm. Trong Hình 9, chúng tôi hiển thị các khung mà mô hình đánh dấu các khung này cách nhau một khoảng thời gian (khoảng một năm). Lượng băng bao phủ giảm rõ rệt trong những năm qua.

# A.5. Video Galleries

Chúng tôi cũng cung cấp phòng trưng bày video với các thay đổi hình ảnh sau:

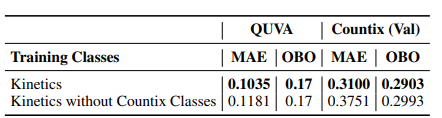
1. PCA 1D của các lần nhúng trong thời gian (liên kết đến video)

2. Ma trận tự tương tự theo thời gian đã học (TSM) của các video khác nhau (liên kết đến video)

# B. Bãi bỏ

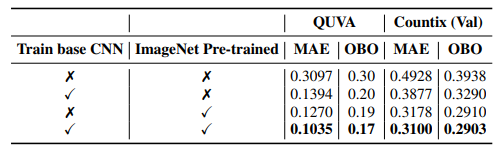
## B.1. Xóa các lớp Countix khỏi Dữ liệu Tổng hợp

Chúng tôi cho thấy rằng việc loại bỏ các lớp được sử dụng trong Countix khỏi nhóm các lớp được sử dụng để tạo dữ liệu tổng hợp có tác động nhỏ đến hiệu suất (xem Bảng 8). Điều này cho thấy rằng việc tổng quát hóa mô hình RepNet không yêu cầu sự hiện diện của các lớp Countix trong tập dữ liệu tổng hợp làm sáng tỏ khía cạnh bất khả tri của lớp trong mô hình của chúng tôi.



Bảng 8: Hiệu quả của việc loại bỏ các lớp Countix khỏi dữ liệu nhập tàu tổng hợp là biên

## B.2. ImageNet Pre-training

Chúng tôi đánh giá tầm quan trọng của đào tạo trước ImageNet đối với mô hình RepNet và báo cáo kết quả trong Bảng 9. Chúng tôi nhận thấy rằng nếu chúng tôi đào tạo mô hình hoàn toàn từ đầu (Hàng 2), chúng ta sẽ đạt được hiệu suất kém hơn 4% so với đào tạo trước (Hàng 4 ). Hiệu suất này vẫn vượt quá các phương pháp hiện đại nhất trong việc đếm số lần lặp lại. Ngoài ra, việc khởi tạo ImageNet của bộ mã hóa mà không cần đào tạo thêm (Hàng 3) là đủ tốt cho nhiệm vụ đếm lặp lại do các mô-đun tiếp theo (TSM và máy biến áp). 

Bảng 9: Lược bỏ đào tạo trước với ImageNet và đào tạo mạng cơ sở hay không. Đối với tất cả các thử nghiệm, chúng tôi đào tạo mô-đun dự đoán chu kỳ và chuyển đổi 3D

## B.3. Tăng cường chuyển động của máy ảnh

Chúng tôi sử dụng các kỹ thuật nâng cao chuyển động của máy ảnh khác nhau cho các video lặp lại tổng hợp như được mô tả trong Hình 5 trong bài báo chính. Chúng tôi cho thấy hiệu quả của việc bỏ qua các phần tăng cường dữ liệu khác nhau. Mỗi phương pháp này dẫn đến lỗi OBO nặng hơn khoảng 1,5% đến 2% và lỗi MAE nặng hơn khoảng 13% đến 26%. Dựa trên các thử nghiệm này, chúng tôi sử dụng tất cả các kỹ thuật nâng cao này cho phần còn lại của các thử nghiệm của chúng tôi.

## B 4. Thay đổi số khung

Trong Bảng 11, chúng tôi báo cáo kết quả khi chúng tôi thay đổi số lượng khung hình mà RepNet lấy làm đầu vào và nhận thấy rằng N = 64 khung hình cung cấp cho chúng tôi hiệu suất tốt nhất. Chúng tôi sử dụng cài đặt này cho tất cả các thí nghiệm trong bài báo chính.

## B.5. Other Architectural Choices

Chúng tôi cũng thay đổi các lựa chọn kiến trúc nhất định được thực hiện trong khi thiết kế RepNet nhưng nhận thấy chúng có tác động nhỏ đến hiệu suất tổng thể.

# C. Implementation Details

## C.1. Detailed Architecture

Trong Bảng 13, chúng tôi trình bày phiên bản chi tiết của kiến trúc RepNet.

## C.2. Kiến trúc của các đường cơ sở thay thế

Đường cơ sở CNN 2D. CNN 2D của chúng tôi bao gồm các lớp tích hợp tiếp theo [32, 64, 128, 256, 512] mỗi lớp có kích thước bộ lọc là 3 × 3. Sau mỗi lớp tích chập, có một hoạt động tổng hợp tối đa của các kích thước 2 × 2 với sải bước 2. Việc tổng hợp trung bình theo không gian toàn cầu được thực hiện trên bản đồ đối tượng địa lý cuối cùng được sử dụng để phân loại độ dài chu kỳ của toàn bộ clip. Chúng tôi cũng đã thử nghiệm với kiến trúc ResNet50 và nhận được hiệu suất tương tự.

LSTM. Chúng tôi sử dụng LSTM tiêu chuẩn được triển khai trong thư viện Tensorflow Keras với 512 đơn vị.

CNN tạm thời 1D. Chúng tôi sử dụng 7 lớp co giật thời gian với tỷ lệ giãn nở [1, 2, 4, 8, 16, 32, 64]. Mỗi lớp chập có kích thước 512 và có kích thước hạt nhân là 2

và có chuẩn hóa hàng loạt. Chúng tôi sử dụng kết nối bỏ qua với

lượng dư cho mỗi lớp.

## C.3. Kết hợp độ dài chu kỳ và chu kỳ ra đặt trong quá trình suy luận

Mô hình của chúng tôi có thể được sử dụng để cùng phát hiện các biến cố định kỳ trong video và chỉ đếm số lần lặp lại trong các phân đoạn lặp lại. Để làm như vậy, chúng tôi lấy mẫu số lần thắng liên tiếp của N khung hình và cung cấp nó làm đầu vào cho RepNet để xuất ra số pi và độ dài chu kỳ trên mỗi khung hình. Chúng tôi xác định số lượng mỗi khung hình là ci = 1/li nếu pi > T khác 0, trong đó T là ngưỡng được chọn cho bộ phân loại. Số lượng video là tổng của tất cả số lượng trên mỗi khung hình: .

# D. Dataset Details

## D.1. Countix Details

Danh sách các lớp được chọn để thu thập dữ liệu trong khi tạo tập dữ liệu Countix được đề cập trong Bảng 14