Hình dạng và vật liệu từ Sound

Zhoutong Zhang Qiujia Li Zhengjia Huang

MIT đại học Cambridge Đại học ShanghaiTech

Jiajun Wu Joshua B. Tenenbaum William T. Freeman

MIT MIT MIT, nghiên cứu của Google

trừu tượng

Nghe một đối tượng rơi xuống đất, con người có thể phục hồi thông tin bao gồm hình dạng của nó thô, vật liệu, và chiều cao rơi xuống. Trong bài báo này, chúng tôi xây dựng máy để xấp xỉ năng lực như vậy. Đầu tiên chúng ta bắt chước kiến thức của con người về thế giới vật chất bằng cách xây dựng một hiệu quả. Sau đó, chúng tôi trình bày một cách tiếp cận phân tích tổng hợp để suy ra tính chất của các đối tượng rơi xuống. Chúng tôi tiếp tục đẩy mạnh quá trình này bằng cách học một ánh xạ từ một làn sóng âm thanh phản anh tính chất đối tượng. và sử dụng các giá trị dự đoán để khởi tạo suy luận. Ánh xạ này này có thể được xem như là một xấp xỉ của tri giác con người học được từ kinh nghiệm quá khứ. mô hình của chúng tôi hoạt động tốt trên cả hai clip âm thanh và ghi âm thực tổng hợp mà không đòi hỏi bất kỳ dữ liệu được chú thích. Chúng tôi tiến hành nghiên cứu hành vi để so sánh phản ứng của con người với chúng ta về ước hình dạng đối tượng, vật liệu, và rơi chiều cao từ âm thanh. mô hình của chúng tôi đạt được hiệu suất gần như con người.

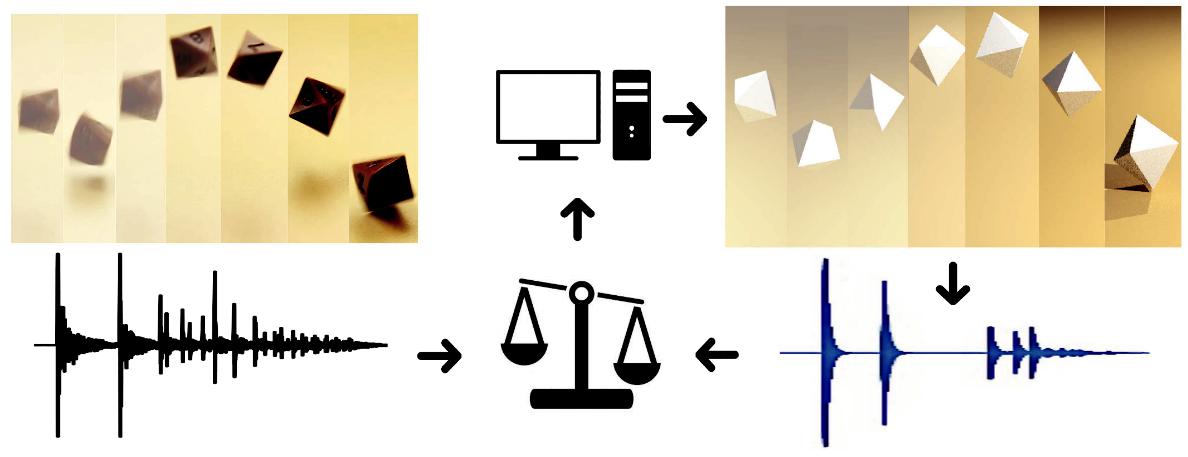
1. Giới thiệu

Từ đoạn clip ngắn của tương tác đối tượng, con người có thể thu hồi số đối tượng có liên quan, cũng như tài liệu của họ và mượt bề mặt [ Zwicker và Fastl , 2013 , Kunkler-Peck và Turvey, 2000 , Siegel et al. , 2014 ]. Làm thế nào để hệ thống nhận thức của chúng tôi thu thập rất nhiều nội dung từ 1 vài thí nghiệm? Vai trò gì của kinh nghiệm quá khứ trong việc tìm hiểu dữ liệu thính giác?

Đối với hiểu biết về cảnh vật từ đầu vào trực quan, nghiên cứu hành vi và tính toán gần đây cho thấy rằng trực giác con người có thể được giải thích như là mô phỏng gần đúng, xác suất của một động cơ vật lý tinh thần [ Battaglia et al. , 2013 , Sanborn et al. , 2013 ]. Những nghiên cứu này cho thấy rằng não mã hóa phong phú, kiến thức về tính chất vật lý của các đối tượng và các luật cơ bản của sự tương tác vật lý giữa các đối tượng. Để hiểu, lý do, và dự đoán về một cảnh vật dường như dựa trên mô phỏng từ động cơ vật lý tinh thần.

Trong bài báo này, chúng tôi phát triển một hệ thống tính toán để giải thích các clip âm thanh của các đối tượng rơi xuống, lấy cảm hứng từ ý tưởng rằng con người có thể sử dụng một động cơ vật lý như là một phần của một mô hình sinh sản để hiểu thế giới vật chất. mô hình sinh sản của chúng tôi có ba thành phần. Đầu tiên là biểu diễn đối tượng bao gồm hình dạng 3D của nó, vị trí trong không gian, và tính chất vật lý như khối lượng, hệ số giảm xóc và đàn hồi. Chúng tôi mong muốn suy ra tất cả những thuộc tính từ đầu vào thính giác.

Thành phần thứ hai là một hiệu quả, động cơ tổng hợp âm thanh dựa trên vật lý. Với một thiết lập cảnh vật ban đầu và các thuộc tính đối tượng, động cơ mô phỏng chuyển động của đối tượng và tạo quỹ đạo của nó bằng cách sử dụng vật cứng. Nó cũng tạo ra thông tin va chạm tương ứng - khi nào, ở đâu và sự va chạm như thế nào. Quỹ đạo của đối tượng và thông tin va chạm sau đó được kết hợp với thống kê âm thanh được tính toán trước để tạo ra âm thanh mà nó tạo ra trong suốt sự kiện va chạm. Với mô hình chuyển tiếp hiệu quả, chúng ta có thể phỏng đoán các thuộc tính đối tượng bằng cách sử dụng phân tích theo tổng hợp; cho mỗi clip âm thanh, chúng tôi muốn tìm kiếm một tập hợp các biến tiềm ẩn phù hợp nhất với tái tạo nó.



Hình 1: Cho một âm thanh của một đối tượng duy nhất rơi xuống, chúng tôi sử dụng mô hình sinh thực của chúng tôi để suy ra các biến tiềm ẩn mà có thể tạo ra âm thanh tốt nhất.

Thành phần thứ 3 là một cái hàm dùng để đo lường khoảng cách nhận thức giữa hai âm thanh. Việc thiết kế một hàm khả năng như vậy thường khó khăn; Tuy nhiên, chúng tôi nhận thấy rằng các tính năng như phổ âm có hiệu quả khi các biến tiềm ẩn có hạn chế về tự do. Điều này thúc đẩy chúng ta suy ra các biến tiềm ẩn thông qua các phương pháp như lấy mẫu Gibbs, nơi chúng tôi tập trung vào gần đúng xác suất có điều kiện của một biến duy nhất cho các biến số khác.

Các suy luận có thể được tăng cường hơn nữa với một mô hình học tập tự giám sát lấy cảm hứng từ các giai đoạn thức / ngủ trong máy Helmholtz [ Dayan et al. , 1995 ]. Chúng tôi đào tạo một mạng lưới thần kinh sâu như mô hình công nhận đặc tính đối tượng hồi quy từ âm thanh, nơi dữ liệu huấn luyện được tạo bằng thuật toán suy luận của chúng tôi. Sau đó, đối với bất kỳ clip âm thanh nào trong tương lai, đầu ra của mô hình nhận dạng có thể được sử dụng làm khởi tạo tốt cho thuật toán lấy mẫu để hội tụ nhanh hơn.

Chúng tôi đánh giá các mô hình của chúng tôi về một loạt các nhiệm vụ nhận thức: suy ra hình dạng vật thể, vật liệu và chiều cao từ âm thanh ban đầu. Chúng tôi cũng thu thập các phản ứng của con người cho từng công việc và so sánh chúng với các ước tính mô hình. Kết quả của chúng tôi cho thấy rằng đầu tiên, con người khá thành công trong những nhiệm vụ này; thứ hai, mô hình của chúng tôi không chỉ phù hợp chặt chẽ với những thành công của con người mà còn tạo ra những lỗi tương tự như con người. Đối với những đánh giá định lượng này, chúng tôi chủ yếu sử dụng dữ liệu tổng hợp, nơi có các nhãn đã có. Chúng tôi tiếp tục đánh giá mô hình trên bản ghi âm để chứng minh rằng nó cũng hoạt động tốt trên các âm thanh trong thế giới thực.

Chúng tôi có ba đóng góp trong bài báo này. Đầu tiên, chúng tôi đề xuất một mô hình mới để ước tính các thuộc tính vật lý từ các âm thanh đầu vào bằng cách kết hợp các phản hồi của một công cụ vật lý và một công cụ âm thanh vào quá trình suy luận. Thứ hai, chúng tôi kết hợp một mạng lưới công nhận sâu sắc với mô hình sinh sản để suy luận hiệu quả hơn. Thứ ba, chúng tôi đánh giá mô hình của chúng tôi và so sánh nó với con người trên nhiều nhiệm vụ đánh giá khác nhau, và chứng minh mối tương quan giữa phản ứng của con người và ước tính mô hình.

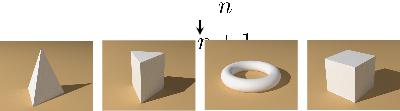
2 Công việc liên quan

**Nhận thức thính giác và thị giác của con người** Các nhà nghiên cứu Psychoacoustics (Dạng tâm lý học liên quan đến âm thanh) đã khám phá cách con người có thể suy ra các đặc tính đối tượng, bao gồm hình dạng, vật liệu và kích thước, từ âm thanh trong những thập kỷ qua [Zwicker và Fastl, 2013, Kunkler-Peck và Turvey, 2000, Rocchesso và Fontana, 2003, Klatzky và cộng sự, 2000, Siegel và cộng sự, 2014]. Gần đây, McDermott et al. [2013] đã đề xuất các biểu diễn âm thanh nhỏ gọn nắm bắt thông tin ngữ nghĩa và là thông tin về nhận thức thính giác của con người.

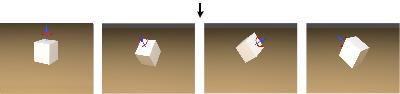
**Mô phỏng âm thanh** Công cụ tổng hợp âm thanh của chúng tôi xây dựng và mở rộng các hệ thống mô phỏng âm thanh hiện có trong đồ họa máy tính và tầm nhìn máy tính [O'Brien et al., 2001, 2002, James et al., 2006, Bonneel et al., 2008, Van den Doel và Pai, 1998, Zhang và cộng sự, 2017]. Van den Doel và Pai [1998] mô phỏng rung động đối tượng bằng phương pháp phần tử hữu hạn và xấp xỉ đối tượng rung như một nguồn điểm đơn. O’Brien et al. [2001, 2002] đã sử dụng phương pháp Rayleigh để ước tính các giải pháp phương trình sóng cho chất lượng tổng hợp tốt hơn. James et al. [2006] đề xuất giải quyết Helmholtz

2

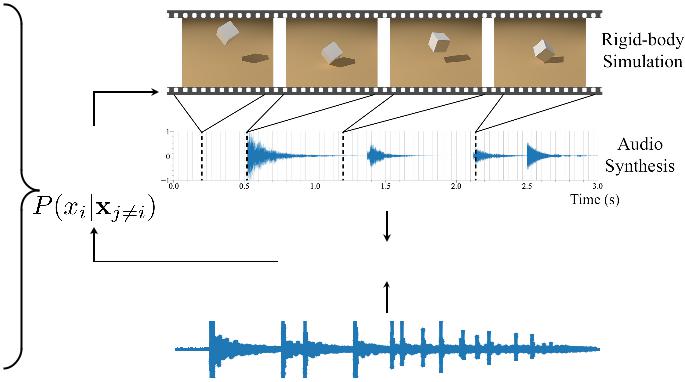
lặp



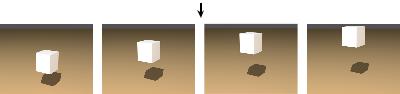
hình dáng



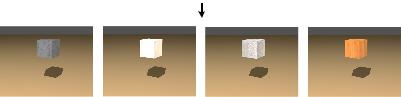
Generative mẫu



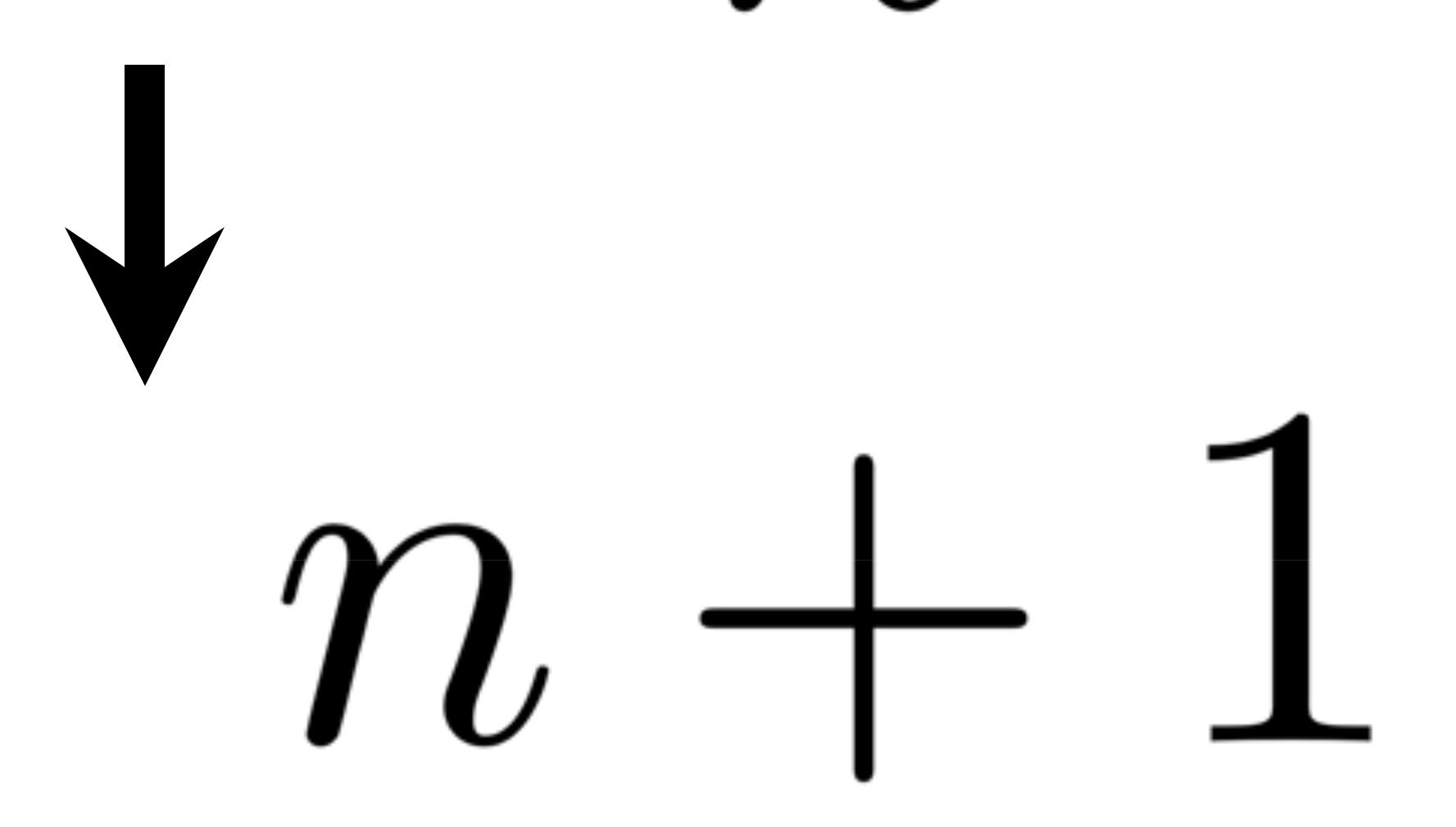
|  |  |
| --- | --- |
| Vòng xoay | Gibbs lấy |
|  | mẫu |
| Chiều cao |  |



Khả năng Chức năng



vật chất



lặp target âm thanh

Hình 2: Đường dẫn suy luận của chúng ta. Chúng tôi sử dụng lấy mẫu Gibbs trên các biến tiềm ẩn. Xác suất có điều kiện là xấp xỉ bằng cách sử dụng tỉ lệ giống nhau giữa âm thanh được tái tạo và âm thanh đầu vào.

phương trình sử dụng phương pháp phần tử ranh giới, nơi chế độ rung của từng đối tượng được xấp xỉ bằng một tập hợp các điểm rung. Gần đây, Zhang et al. [2017] đã xây dựng một framework để tổng hợp dữ liệu nghe nhìn lớn. Trong bài báo này, chúng tôi chạy framework của Zhang et al. [2017] để đạt được gần thời gian thực sự biểu diễn, và khám phá đại diện đối tượng học tập từ âm thanh với các công cụ tổng hợp trong vòng lặp.

**Nhận thức đối tượng vật lý** Đã có sự quan tâm ngày càng tăng trong việc hiểu các đặc tính đồ vật vật lý, như khối lượng và ma sát, từ động lực đầu vào hoặc cảnh động [Chang et al., 2017, Battaglia và cộng sự, 2016, Wu và cộng sự, 2015, 2016, 2017]. Phần lớn các nghiên cứu hiện có đã tập trung vào việc suy ra các thuộc tính đối tượng từ dữ liệu trực quan. Gần đây, các nhà nghiên cứu đã bắt đầu khám phá các biểu diễn đối tượng học tập từ âm thanh. Owens et al. [2016a] đã cố gắng suy ra các thuộc tính vật chất từ ​​âm thanh, tập trung vào kịch bản đập đối tượng bằng một dùi trống. Owens et al. [2016b] tiếp tục chứng minh tín hiệu âm thanh có thể được sử dụng như giám sát về các khái niệm đối tượng học tập từ dữ liệu trực quan, và Aytar et al. [2016] đề xuất tìm hiểu các biểu diễn âm thanh từ các khung hình video tương ứng. Zhang et al. [2017] đã thảo luận về vai trò bổ sung của dữ liệu thính giác và thị giác trong việc phục hồi cả thuộc tính đối tượng hình học và vật lý. Trong bài báo này, chúng ta học các biểu diễn vật thể vật lý từ âm thanh thông qua sự kết hợp các mô hình nhận dạng sâu mạnh mẽ và các phương pháp suy luận phân tích tổng hợp.

**theo phân tích tổng hợp** Framework của chúng tôi cũng liên quan đến lĩnh vực phân tích tổng hợp, hoặc các mô hình sinh sản với các đề xuất dựa trên dữ liệu [Yuille và Kersten, 2006, Zhu và Mumford, 2007, Wu và cộng sự, 2015], như chúng tôi đang tích hợp công cụ đồ họa làm bộ tổng hợp hộp đen. Không giống như các phương pháp trước đó tập trung chủ yếu vào việc giải thích dữ liệu trực quan, công việc của chúng tôi nhằm mục đích suy ra các tham số tiềm ẩn từ dữ liệu thính giác. Vui lòng tham khảo Bever and Poeppel [2010] để xem xét các phương pháp phân tích theo tổng hợp.

**3 năng lực, Công cụ âm thanh dựa trên vật lý**

Dựa theo suy luận của chúng tôi là một công cụ tổng hợp âm thanh hiệu quả. Trong phần này, trước tiên chúng tôi đưa ra một tổng quan ngắn gọn về các công cụ tổng hợp hiện có và sau đó trình bày các cải tiến kỹ thuật của chúng tôi về việc tăng tốc chúng để hiển thị thời gian thực trong thuật toán suy luận của chúng tôi.

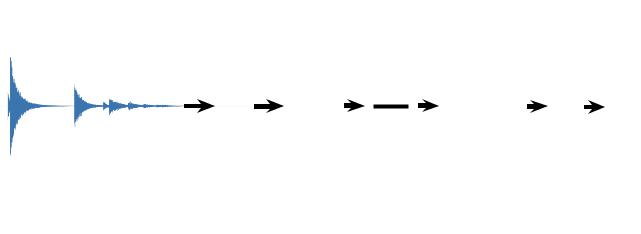
**3.1 Engine âm thanh tổng hợp**

. Thứ nhất, mô phỏng vật cứng

Các Engine tổng hợp âm thanh tạo ra âm thanh trung thực bằng cách mô phỏng vật lý tạo ra sự tương tác giữa một đối tượng và môi trường, trong đó các định luật Newton quyết định chuyển động và va chạm của đối tượng theo thời gian. Mỗi va chạm làm cho đối tượng tạo độ rung trong các mẫu nhất định, thay đổi áp suất không khí xung quanh bề mặt. Những độ rung này lan truyền trong không khí tới đầu ghi và tạo ra âm thanh của quá trình vật lý này.

3

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  | Cài đặt | Thời gian ( *S)* |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  | thuật toán ban đầu | 30,4 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  | cắt biên độ | 24,5 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  | chế độ chính | 12,7 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  | Multi-threading | 1,5 |  |
| dạng sóng |  |  |  |  | SoundNet-8 | | |  |  |  |  |  |  |
|  |  |  |  |  | | | | |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | Tất cả các | 0.8 |  |
|  | pool1 conv1 | | | |  |  |  |  | | | | |  |  |
| âm thanh sóng |  |  | conv7 pool7 fc ...... | | | | | |  |



Hình 3: mạng chập sâu 1D của chúng tôi. kiến trúc của nó sau đó trong Aytar et al. [ 2016 ], Nơi sóng âm thanh thô được chuyển tiếp qua lớp của nhóm chuyển tiếp và sau đó được chuyển đến một lớp được kết nối đầy đủ để tạo ra đầu ra.

Bảng 1: Tăng tốc phân chia từng kỹ thuật mà chúng tôi đã áp dụng. Thời gian được đánh giá bằng cách tổng hợp một âm thanh với 200 va chạm. Hàng cuối cùng báo cáo thời gian cuối cùng sau khi áp dụng tất cả các kỹ thuật.

**Mô phỏng vật cứng** Với vị trí và góc quay 3D của đối tượng, khối lượng và đàn hồi của nó, một công cụ vật lý có thể mô phỏng các quá trình vật lý và xuất ra thông tin vị trí, góc quay và va chạm của đối tượng theo thời gian. Việc triển khai của chúng tôi sử dụng một mã nguồn mở engine vật lý, Bullet [Coumans, 2010]. Chúng tôi sử dụng một bước thời gian 1/300 giây để đảm bảo tính chính xác của mô phỏng. Tại mỗi bước, chúng ta ghi lại tư thế 3D và vị trí của vật thể, cũng như vị trí, cường độ và hướng va chạm. Âm thanh do vật thể tạo ra sau đó có thể được xấp xỉ bằng cách tích lũy âm thanh gây ra bởi những va chạm xung đột rời rạc trên bề mặt của nó.

**Tổng hợp âm thanh** Quy trình tổng hợp âm thanh được xây dựng dựa trên công việc trước đây về mô phỏng âm thanh thực tế [James et al., 2006, Bonneel và cộng sự, 2008, O’Brien và cộng sự, 2001]. Để tạo điều kiện tổng hợp nhanh, quá trình này được phân tách thành hai mô-đun, một mô-đun ngoại tuyến và một mô-đun trực tuyến. Phần ngoại tuyến đầu tiên sử dụng các phương thức phần tử hữu hạn (FEM) để thu được các chế độ rung của đối tượng, dựa vào hình dạng và mô dun đàn hồi của đối tượng. Các chế độ rung này sau đó được sử dụng như các điều kiện biên Neumann của phương trình Helmholtz, có thể được giải quyết bằng phương pháp phần tử biên (BEM). Chúng tôi sử dụng các kỹ thuật được đề xuất bởi James et al. [2006] để ước tính giải pháp bằng cách mô hình hóa các trường áp suất với một tập hợp các điểm rung nhỏ. Lưu ý rằng việc tính toán ở trên chỉ phụ thuộc vào các thuộc tính bên trong của đối tượng chẳng hạn như hình dạng và đàn hồi, nhưng không phụ thuộc vào bên ngoài như vị trí và vận tốc của vật thể. Điều này cho phép chúng tôi tính toán trước một số lượng các cấu hình hình dạng-mô-đun trước khi mô phỏng; chỉ cần tính toán tối thiểu trong mô phỏng trực tuyến.

Phần trực tuyến của công cụ âm thanh tải các xấp xỉ được tính toán trước và phân hủy các xung trên bề mặt lưới của đối tượng thành các cơ sở phương thức của nó. Tại điểm quan sát, engine đo các thay đổi áp suất gây ra bởi các rung động trong mỗi trường hợp, và tổng hợp chúng để tạo ra âm thanh mô phỏng. Một đánh giá về độ trung thực của các mô phỏng này có thể được tìm thấy trong Zhang et al. [2017].

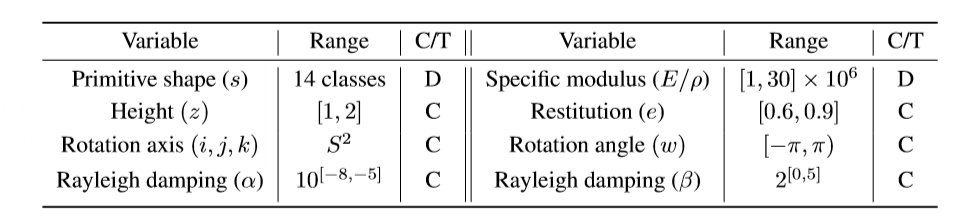
**3.2. Tăng tốc âm thanh tổng hợp**

Phân tích suy luận tổng hợp yêu cầu engine âm thanh phải hiệu quả cao; tuy nhiên, việc thực hiện đơn giản trên quy trình mô phỏng ở trên sẽ tốn kém về mặt tính toán. Do đó, chúng tôi trình bày các cải tiến kỹ thuật để đẩy nhanh tính toán đến gần thời gian thực.

Đầu tiên, chúng tôi chọn các trường hợp quan trọng nhất được phù hợp bởi mỗi xung cho đến khi tổng năng lượng của chúng tôi đạt tới 90% năng lượng của xung. Bỏ qua các thành phần âm thanh được tạo ra bởi các trường hợp ít quan trọng hơn làm giảm thời gian tính toán khoảng 50%. Thứ hai, chúng tôi ngừng quá trình tổng hợp nếu biên độ của âm thanh bị giảm đi dưới một ngưỡng nhất định, vì nó không có khả năng nghe được. Thứ ba, chúng tôi song song quá trình tổng hợp bằng cách giải quyết các va chạm riêng biệt, để mỗi có thể tính toán trên một tiến trình độc lập. Sau đó, chúng tôi tích hợp chúng vào bộ đệm chia sẻ để tạo âm thanh cuối cùng theo mốc thời gian của chúng. Kết quả được thể hiện trong Bảng 1. Tổng hợp âm thanh trực tuyến chỉ chứa các biến được tách hoàn toàn khỏi giai đoạn ngoại tuyến, cho phép chúng ta tự do thao tác các biến khác với chi phí tính toán nhỏ trong khi mô phỏng.

**3.3 Sinh ra Yếu tố kích thích**

Bởi vì các bản ghi âm thực sự với các nhãn phong phú khó có được, chúng tôi tổng hợp các clip âm thanh ngẫu nhiên bằng cách sử dụng mô phỏng dựa trên vật lý để đánh giá các mô hình của chúng tôi.



Bảng 2: Các biến trong mô hình sinh sản của chúng tôi, nơi cột C / T cho biết mẫu diễn ra trong liên tục (C) hoặc rời rạc (D) miền, và các giá trị bên trong dấu ngoặc là phạm vi chúng tôi thống nhất lấy mẫu từ. Xoay được xác định trong quaternions.

Cụ thể chúng tôi tập trung vào một kịch bản - hình dạng nguyên thủy rơi xuống đất. Trước tiên, chúng tôi xây dựng tập dữ liệu âm thanh bao gồm 14 tập dữ liệu nguyên thủy (một số được hiển thị trong Bảng 2), mỗi loại có 10 mô đun cụ thể khác nhau (được định nghĩa là mức độ đàn hồi). Sau khi tính toán trước thiệt lập từng không gian của chúng, chúng ta có thể tạo ra các clip âm thanh tổng hợp trong một thời gian gần như thời gian thực. Bởi vì quá trình đối tượng rơi xuống mặt đất tương đối nhanh, chúng tôi thiết lập tổng thời gian mô phỏng của từng kịch bản thành 3 giây. Chi tiết về thiết lập của chúng tôi có thể được tìm thấy trong Bảng 2.

**4 Suy luận**

Trong phần này, chúng tôi điều tra bốn mô hình để suy luận về tính chất đối tượng, mỗi tương ứng với một điều kiện đào tạo khác nhau. Lấy cảm hứng từ cách con người có thể suy ra thông tin cảnh sử dụng một engine vật lý kim loại[ Battaglia et al. , 2013 , Sanborn et al. , 2013 ], Chúng tôi bắt đầu từ một mô hình không có giám sát nơi đầu vào là chỉ có một trường hợp thử nghiệm duy nhất không có chú thích. Chúng tôi áp dụng lấy mẫu Gibbs qua các biến tiềm ẩn để tìm sự kết hợp tái tạo tốt nhất âm thanh đã cho.

Sau đó chúng tôi mở rộng mô hình để bao gồm một mạng lưới thần kinh sâu, tương tự như những gì con người có thể học hỏi từ kinh nghiệm quá khứ của họ. Mạng lưới này được huấn luyện sử dụng nhãn suy ra từ mô hình không có giám sát. Trong suy luận, các thuật toán lấy mẫu sử dụng mạng lưới dự đoán khởi tạo. học tập mô hình tự giám sát này tăng tốc lên hội tụ.

Chúng tôi cũng điều tra một trường hợp thứ ba, khi nhãn có thể được cung cấp nhưng cực kỳ thô. Chúng tôi đầu tiên đào tạo một mô hình công nhận với nhãn yếu, sau đó chọn một cách ngẫu nhiên một trường hợp đến từ những nhãn đó như một khởi tạo cho kết luận phân tích theo tổng hợp của chúng tôi.

Cuối cùng, để hiểu giới hạn hiệu suất, chúng tôi đào tạo một mạng lưới thần kinh sâu sắc với dữ liệu đầy đủ nhãn mà mang lại hiệu suất vượt giới hạn.

4.1 **Các mô hình**

**Có giám sát** Với một clip âm thanh *S* đã cho*,* chúng tôi muốn thu thập các biến tiềm ẩn x để làm cho tái tạo âm thanh với *g(* x) tạo ra âm thanh tương tự như *S.* Cho *L ( ·, ·)* là một hàm khả năng đo tương đồng giữa hai âm thanh, sau đó mục tiêu củachúng tôi là để tối ưu hóa *L (g (* x), *S).* chúng tôi biểu thị *L (g (* x), *S)* như *p (* x) cho ngắn gọn. Để tìm x nhằm tối đa hóa *p(* x), *p (* x) có thể được coi là một phân phối ˆ*p (* x) tối ưu lại bởi một hàm phân vùng chưa biết *Z.* Vì chúng ta không có một hình thức chính xác cho *p (.),* cũng không ˆ*p(*x), chúng tôi áp dụng Gibbs lấy mẫu để vẽ mẫu từ *p(* x). Cụ thể, tại vòng quét *t,* chúng tôi cập nhật mỗi biến x *tôi* bằng cách vẽ mẫu từ



xác suất có điều kiện như vậy là đơn giản để gần đúng. Ví dụ, để lấy mẫu mô đun đàn hồi trên các biến khác, chúng ta có thể sử dụng âm phổ như một tính năng và đo lường *l*2khoảng cách giữa các âm phổ của hai âm thanh, bởi vì mô đun đàn hồi chỉ sẽ ảnh hưởng đến tần số tại mỗi vụ va chạm. Thậtvậy, chúng ta có thể sử dụng âm phổ như các tính năng cho tất cả các giá trị ngoại trừ chiều cao. Kể từ khi chiều cao có thể được suy ra từ thời điểm va chạm đầu tiên, một hàm khả năng đơn giản có thể được thiết kế như đo lường sự chênh lệch thời gian giữa tác động đầu tiên trong hai âm thanh. Lưu ý rằng đây chỉ là một biện pháp gần đúng: hình dạng và góc quay đối tượng cũng ảnh hưởng, mặc dù chỉ một chút, thời điểm va chạm đầu tiên.

Để lấy mẫu từ các xác suất có điều kiện, chúng tôi áp dụng các thuật toán Metropolis-Hastings, nơi mẫu được rút ra từ một phân phối Gaussian và được chấp nhận bằng cách lật một đồng xu theo khả năng của nó để so tương đồng với các mẫu trước đó. Đặc biệt, chúng tôi tính toán l2 khoảng cách dt trong không gian đặc trưng giữa g(xt+1) và S. Đối với một mẫu mới xt+1, chúng tôi cũng tính toán l2 khoảng cách dt + 1 trong không gian đặc trưng giữa g(xt + 1) và S. Các mẫu mới được chấp nhận nếu dt + 1nhỏ hơn hoặc bằng dt*;* nếu không thì, x*t +* 1được chấp nhận với xác suất exp(*-(dt+*1*-dt)/T) ,* Nơi *T* là một thời gian khác nhau lấy cảm hứng từ thuậttoán mô phỏng kim loại. Trong việc thực hiện của chúng tôi, *T* được thiết lập như một hàm bậc hai của số quét MCMC hiện tại *t.*

**Học tự giám sát** Để đẩy nhanh quá trình lấy mẫu ở trên, chúng tôi đề xuất một mô hình tự giám sát, đó là tương tự như một máy Helmholtz được đào tạo bởi các thuật toán thức/ ngủ. Chúng tôi đầu tiên đào tạo một mạng lưới thần kinh sâu, có các nhãn được tạo ra bởi các mô hình suy luận không có giám sát đề nghị ở trên cho một số giới hạn lặp lại. Đối với một clip âm thanh mới, mô hình tự giám sát của chúng tôi sử dụng kết quả từ mạng thần kinh như khởi tạo, và sau đó chạy thuật toán phân tích theo tổng hợp của chúng tôi để lại suy luận. Bằng cách sử dụng những kinh nghiệm trong quá khứ mà đào tạo mạng lưới, quá trình lấy mẫu bắt đầu từ một vị trí tốt hơn và đòi hỏi lặp đi lặp lại ít hơn để tạo ra mô hình không có giám sát

**Học yếu giám sát** Chúng tôi tiếp tục điều tra các trường hợp giám sát yếu có thể hữu ích cho việc thúc đẩy quá trình suy luận. Bởi vì các biến tiềm ẩn chúng tôi hướng đến thu thập rất khó để có được trong cài đặt thế giới thực, nó thực tế hơn khi cho rằng chúng ta có thể có được nhãn rất thô, chẳng hạn như các loại vật liệu, các thuộc tính sơ bộ hình dạng của đối tượng, chiều cao độ rơi, *vv* Dựa trên những giả định như vậy, chúng tôi giảm nhãn thật cho tất cả các giá trị. Đối với hình dạng nguyên thủy, ba thuộc tính này là định nghĩa, cụ thể là “cạnh”, “bề mặt cong,” và “nhọn”. Đối với thông số vật liệu, ví dụmô đun cụ thể, giảm xóc và trở về hình dáng ban đầu Rayleigh, họ được ánh xạ tới thép, gốm, nhựa và gỗ bằng cacsht tìm các tương quan gần nhất với những thông số vật liệu thực sự. Chiều cao được chia thành “thấp” và “cao”. Một mạng lưới thần kinh xoắn sâu được đào tạo về dữ liệu tổng hợp của chúng tôi với nhãn thô. Như thể hiện trong hình 4 , Mặc dù được đào tạo sử dụng nhãn thô, mạng của chúng tôi học được tính năng rất giống với những người học bằng cách mạng lưới hoàn toàn có giám sát. Để vượt qua nhãn thô, mô hình không có giám sát được áp dụng bằng cách sử dụng khởi tạo được đề xuất bởi các mạng thần kinh.

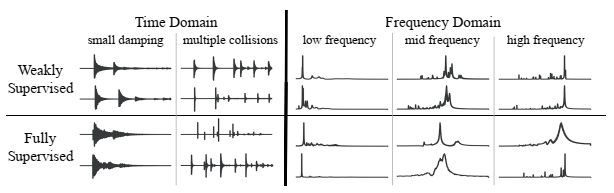
**Học giám sát đầy đủ** Để khám phá giới hạn hiệu suất trong những nhiệm vụ suy luận, chúng tôi đào tạo một mô hình oracle với nhãn thực địa. Để hình dung trừu tượng và các tính năng đặc trưng học bằng mô hình oracle, ta vẽ đồ thị đầu vào mà tối đa kích hoạt một số đơn vị ẩn trong lớp cuối cùng của mạng. Số liệu 4 minh họa một số các dạng sóng hấp dẫn nhất. Một lựa chọn trong số họ đã học cách nhận ra mô hình thời gian cụ thể, và những thứ khác là nhạy cảm với tần số đặc hiệu. mô hình tương tự cũng được tìm thấy trong các mô hình học yếu và đầy đủ giám sát.

**4.2 Hiệu suất mô hình tương phản**

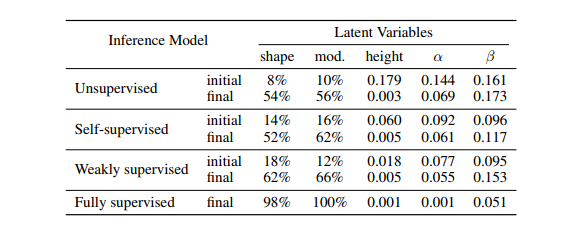
Chúng tôi đánh giá như thế nào mô hình của chúng tôi thực hiện dưới thiết lập khác nhau, nghiên cứu kinh nghiệm quá khứ hoặc nhãn thô có thể cải thiện kết quả không có giám sát như thế nào. Đầu tiên, chúng tôi trình bày các chi tiết thực hiện của cả bốn mô hình, sau đó so sánh kết quả của họ trên tất cả các task suy luận.

**Cài đặt lấy mẫu** Chúng tôi thực hiện 80 lần quét lấy mẫu MCMC trên tất cả các 7 Các biến tiềm ẩn; cho mỗi quét, mỗi biến được lấy mẫu hai lần. Shape, mô đun cụ thể và góc quay được lấy mẫu bởi sự phân bố đồng đều giữa các kích thước tương ứng của nó. Đối với các biến liên tục khác, chúng ta xác định một biến Gaussian phụ trợ *x*­i *~ N(μ*­i*,* σi2*)* để lấy mẫu, nơi giá trị trung bình *μi* được dựa trên tình trạng hiện thời. Để đánh giá hàm tương đồng giữa đầu âm thanh vào và mẫu (cả với tỷ lệ mẫu của 44.1k), chúng tôi tính toán âm phổ của tín hiệu sử dụng một cửa sổ Tukey có chiều dài 5000 với một sự 2.000 mẫu chồng chéo. Đối với mỗi cửa sổ, một điểm 10.000 biến đổi Fourier được áp dụng.

**Cài đặt Learning sâu** mô hình công nhận hoàn toàn giám sát và tự giám sát của chúng tôi sử dụng Tecture archi- của SoundNet-8 [Aytar et al. , 2016 ] Như Hình 3, Trong đó có một làn sóng âm thanh thô tùy tiện miễn là một đầu vào, và tạo ra một vector đặc trưng 1024 dim. Chúng tôi thêm vào đó một lớp đầy đủ kết nối để tạo ra một vector 28-dim như đầu ra mạng lưới thần kinh cuối cùng. Các 14 dimemsion đầu tiên là mã hóa một nóng của hình dạng nguyên thủy và 10 dimemsion tiếp theo là mã hóa của các mô đun đặc trưng. 4 dimemsion cuối cùng thoái chiều cao ban đầu, hai hệ giảm xóc Rayleigh và



Hình 4: Trực quan của hai sóng âm thanh hàng đầu mà kích hoạt các đơn vị ẩn ý nghĩa nhất, trong phạm vi thời gian và âm phổ. đặc điểm chung của nó có thể ánh xạ các giá trị của một số biến tiềm ẩn, *ví dụ* giảm xốc Rayleigh, đàn hời và mô đun cụ thể từ trái sang phải. Cả hai mô hình một cách yếu ớt và đầy đủ giám sát chụp tính năng tương tự.

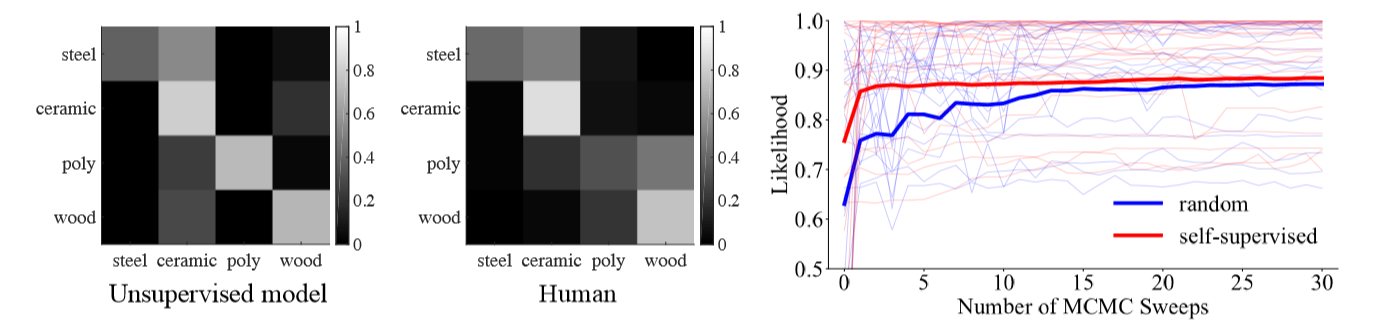


Bảng 3: phân loại độ chính xác ban đầu và cuối cùng (như tỷ lệ phần trăm) và tham số MSE lỗi của ba mô hình suy luận khác nhau sau 80 lần lặp của MCMC. Không có giám sát ban đầu cho thấy hiệu suất ngẫu nhiên. Kết quả từ mô hình giám sát đầy đủ cho ra giới hạn hiệu suất. *α* và *β* được hệ số giảm xóc Rayleigh.

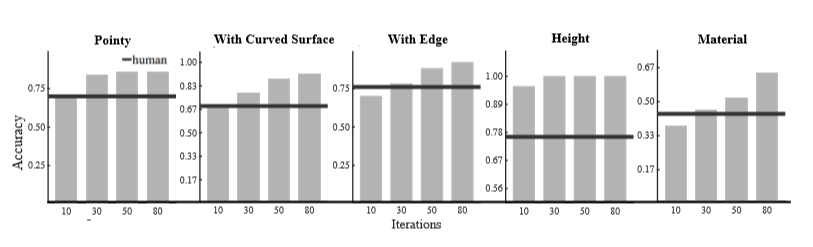
trở về hình dạng ban đầu tương ứng. Tất cả các kích thước hồi quy được chuẩn hóa đến một phạm vi [ *-* 1, 1] . Mô hình một cách yếu ớt giám sát bảo tồn cấu trúc của toàn giám sát một, nhưng với sản lượng fi NAL 8-dim: 3 cho các thuộc tính hình dạng, 1 cho chiều cao, và 4 đối với nguyên liệu. Chúng tôi sử dụng ngẫu nhiên gradient descent cho đào tạo, với tỷ lệ học của 0,001, một đà 0,9 và kích thước lô 16. Mean Square Error (MSE) mất được sử dụng để lan truyền ngược. Chúng tôi thực hiện khuôn khổ của chúng tôi trong Torch7 [ Collobert et al. , 2011 ], Và được đào tạo tất cả các mẫu từ đầu.

**Kết quả** những kết quả cho bốn mô hình suy luận đề xuất trên được thể hiện trong Bảng 3. Đối với hình dạng và mô đun đặc trưng, chúng tôi đánh giá các kết quả như độ chính xác phân loại; cho chiều cao, hệ số giảm xóc Rayleigh, và trở về nguyên dạng, kết quả được đánh giá bởi MSE. Trước khi tính MSE, chúng ta đơn giản hóa giá trị của mỗi biến tiềm ẩn trong khoảng [ *-* 1, 1] , do đó số điểm MSE có thể so sánh giữa các biến.

Từ Bảng 3 , Chúng ta có thể kết luận rằng các mô hình tự giám sát và yếu giám sát có lợi hơn từ khởi tạo để các thuật toán phân tích tổng hợp, đặc biệt là trên bốn biến ẩn liên tục cuối cùng. Người ta cũng có thể quan sát rằng độ chính xác suy luận cuối cùng và MSE đã hoành chỉnh hơn so với trường hợp không có giám sát. Để minh họa cho tốc độ hội tụ, ta vẽ đồ thị giá trị tương đồng, exp ( *- kd)* nơi *d* là khoảng cách các âm thanh, cùng lặp lại của MCMC trong hình 5 . Đường cong trung bình của mô hình tự giám sát đáp ứng sự mong đợi của chúng tôi, *ví dụ,* nó hội tụ nhanh hơn nhiều so với mô hình không có giám sát, và đạt đến một khả năng cao hơn một chút ở phần cuối của 30 lần lặp lại. Mô hình hoàn toàn có giám sát, được đào tạo về 200.000 âm thanh với đầy đủ các nhãn, sản lượng kết quả gần như hoàn hảo cho tất cả các biến tiềm ẩn.



Hình 5: Trái và Giữa: ma trận nhẫm lẫn phân loại được thực hiện bởi con người và mô hình không có giám sát. Bên phải: đường cong trung bình giống nhau trên lặp MCMC.



Hình 6: so sánh hiệu suất người và hiệu suất không có giám sát. Đoạn ngang đại diện cho hiệu suất con người cho mỗi task. Thuật toán của chúng tôi phù hợp chặt chẽ hoạt động của con người.

**5. Đánh giá**

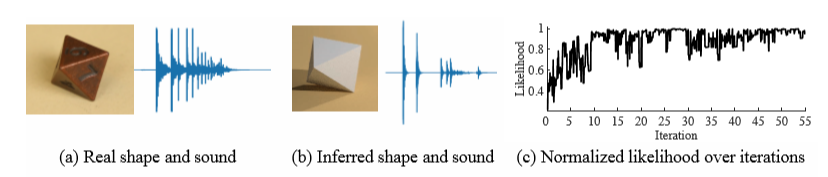
Chúng tôi đánh giá đầu tiên việc thực hiện các thủ tục suy luận của chúng tôi bằng cách so sánh hiệu quả của nó với con người. Việc đánh giá được thực hiện bằng âm thanh tổng hợp với các nhãn thực địa của họ. Sau đó, chúng tôi điều tra xem liệu thuật toán suy luận của chúng tôi hoạt động tốt trên các bản ghi âm thực tế. Với ghi âm thanh, thuật toán của chúng tôi có thể phân biệt các hình dạng từ một tập hợp các ứng cử viên.

**5.1 Nghiên cứu con người**

Chúng tôi tìm kiếm để đánh giá hiệu suất mô hình với con người. Chúng tôi tạo ba nhiệm vụ cho các đối tượng của chúng tôi: Suy luận hình dạng, vật liệu và chiều cao rơi của đối tượng từ âm thanh, các thuộc tính trực quan khi nghe một đối tượng rớt. Những nhiệm vụ được thiết kế trong việc phân loại, nơi mà các nhãn là phù hợp với nhãn thô được sử dụng bởi mô hình yếu giám sát. Nghiên cứu được tiến hành trên Amazon Mechanical Turk. Đối với mỗi thí nghiệm (hình dáng, vật liệu, chiều cao), chúng tôi lựa chọn ngẫu nhiên 52 trường hợp thử nghiệm. Trước khi trả lời câu hỏi kiểm tra, đối tượng được hiển thị 4 ví dụ đào tạo với thực địa như làm quen của các thiết lập. Chúng tôi thu thập được 192 câu trả lời cho các thí nghiệm trên hình dạng suy luận, 566 cho vật liệu, và 492 cho chiều cao, dẫn đến tổng cộng 1.250 câu trả lời.

**Suy luận hình dáng** Sau khi trở thành quen với thí nghiệm, người tham gia được yêu cầu làm nhị phân về hình dạng bằng cách lắng nghe clip âm thanh tổng hợp của chúng tôi. ví dụ trước khi được đưa ra cho mọi người hiểu được sự phân biệt của “mép”, “bề mặt cong,” và “nhọn” thuộc tính. Như thể hiện trong hình 6 , Con người là tương đối tốt tại nhận hình dạng thuộc tính từ âm thanh và khoảng cùng một mức độ năng lực khi các thuật toán không giám sát chạy lặp lại trong 10 *~* 30 lần.

**Suy luận chất liệu** Chúng tôi lấy mẫu các đoạn âm thanh có tính chất vật lý - mật độ, đàn hổi và hệ số giảm xóc – nằm trong vùng lân cận của các tham số của thép, gốm, nhựa và gỗ. Những người tham gia được yêu cầu chọn một trong số bốn loại vật liệu có thể. Tuy nhiên, nó vẫn có thể là một thách thức để phân biệt giữa vật liệu, đặc biệt là khi những mẫu có tương tự giảm xóc và mô đun đặc trưng. Thuật toán của chúng tôi thỉnh thoảng lẫn lộn thép với gốm, gốm và nhựa, điều này phù hợp với hiệu suất của con người, như thể hiện trong hình 5 .



(A) hình Real và âm thanh (B) hình dạng suy ra và âm thanh (C) khả năng bình thường hóa trên lặp đi lặp lại

Hình 7: Kết quả của suy luận trên dữ liệu thế giới thực. Buổi ghi hình thử nghiệm được thực hiện bằng cách thả những con xúc xắc kim loại trong (a).

hình dạng suy ra chúng ta và tái tạo âm thanh được thể hiện trong (b). Khả năng trên lặp được vẽ trong (c).

Suy luận Heights Trong nhiệm vụ này, chúng tôi yêu cầu người tham gia để chọn xem đối tượng được giảm từ một vị trí cao hoặc một thấp. Chúng tôi cung cấp video ví dụ và âm thanh để giúp mọi người neo chiều cao tham khảo. Theo thiết lập cảnh của chúng tôi, thời gian touchdown trong hai thái cực của phạm vi độ cao khác nhau bởi 0.2s. Để giải quyết những sai lệch tiềm ẩn mà các thuật toán có thể được tốt hơn trong việc khai thác thời gian rơi xuống, chúng tôi nói một cách rõ ràng con người mà sự im lặng lúc đầu là thông tin. Thứ hai, chúng tôi đảm bảo rằng các ví dụ neo luôn luôn có sẵn trong kiểm tra, trong đó người tham gia luôn có thể so sánh và tham khảo. Thứ ba, người tham gia phải đóng mỗi bài kiểm tra kẹp bằng tay, và do đó có quyền kiểm soát khi các âm thanh bắt đầu. Cuối cùng, chúng tôi thử nghiệm trên hình dạng đối tượng khác nhau. Bởi vì thời điểm va chạm đầu tiên fi là hình dạng phụ thuộc,

5.2 Chuyển sang Real Scenes

Ngoài các dữ liệu tổng hợp, chúng tôi thiết kế các thí nghiệm thế giới thực để kiểm tra mô hình không có giám sát của chúng tôi. Chúng tôi

chọn ba hình dạng ứng cử viên: tứ diện, octahedron, và khối mười hai mặt. Chúng tôi ghi lại âm thanh một octahedron giảm kim loại trên một

bảng và sử dụng mô hình không có giám sát của chúng tôi để phục hồi các biến tiềm ẩn. Bởi vì các kịch bản thế giới thực có thể giới thiệu

các yếu tố phức tạp mà không thể được mô hình chính xác trong mô phỏng của chúng tôi, một tính năng mạnh mẽ hơn và một số liệu là cần

thiết. Đối với mỗi clip âm thanh, chúng tôi sử dụng phân phối năng lượng dương của nó như tính năng, trong đó có nguồn gốc từ phổ. Một

cửa sổ 2.000 mẫu với một chồng chéo 1.500 mẫu được sử dụng để tính toán sự phân bố năng lượng. Sau đó, chúng tôi sử dụng khoảng

cách mover trái đất (EMD) [ Rubner et al. , 2000 ] Như số liệu, mà là một sự lựa chọn tự nhiên để đo khoảng cách giữa các bản phân phối. Kết

quả suy luận được minh họa trong hình 7 . Sử dụng sự phân bố năng lượng với biện pháp EMD khoảng cách, âm thanh tạo của chúng tôi gắn

năng lượng của mình tại các sự kiện va chạm lớn với âm thanh thực, làm giảm đáng kể sự mơ hồ giữa ba hình dạng ứng cử viên. Chúng tôi

cũng cung cấp bình thường hàm likelihood thêm giờ của chúng tôi để hiển thị lấy mẫu của chúng tôi đã hội tụ để tạo ra các mẫu cao có thể

xảy ra.

6 Kết luận

Trong bài báo này, chúng tôi đề xuất một mô hình mới để ước tính chất vật lý của các đối tượng từ đầu vào thính giác, bằng cách kết hợp các thông tin phản hồi của một fi cient động cơ tổng hợp âm thanh ef trong vòng lặp. Chúng tôi demon- strate khả năng tăng tốc suy luận với các mô hình nhận dạng nhanh. Chúng tôi so sánh mô hình dự báo của chúng tôi với câu trả lời của con người trên một loạt các nhiệm vụ đánh giá và chứng minh mối tương quan giữa phản ứng của con người và ước tính mô hình. Chúng tôi cũng thấy rằng mô hình của chúng tôi khái quát một số dữ liệu thực tế.

Lời cảm ơn

Các tác giả xin trân trọng cảm ơn Changxi Zheng, Eitan Grinspun, và Josh H. McDermott cho các cuộc thảo luận hữu ích. Công trình này được hỗ trợ bởi NSF # 1212849 và # 1447476, ONR Muri N00014-16-1-

Năm 2007, Viện Nghiên cứu Toyota, Samsung, Shell, và Trung tâm não, Minds và Máy móc (NSF STC giải thưởng CCF-1.231.216).

9