Tóm Tắt Video Bằng Mạng Học Sâu Nơ-ron

1st Chung Kim Khánh  
Đại học Khoa học tự nhiên (ĐHQG)Hồ Chí Minh, Việt Nam  
ckkhanh19@clc.fitus.edu.vn

2nd Phạm Nguyễn Anh Quốc  
Đại học Khoa học tự nhiên (ĐHQG)Hồ Chí Minh, Việt Nam  
pnaquoc19@clc.fitus.edu.vn

*Tóm lược*—Công nghệ tóm tắt video dùng để tạo ra bản video mới tóm tắt một cách ngắn gọn nhưng diễn đạt đầy đủ thông tin của video ban đầu. Vì thế, ta thực hiện chọn lọc những đoạn video có thông tin quan trọng để sử dụng. Một số phương pháp đã được đưa ra và cải tiến qua nhiều thập kỷ. Hiện tại, các phương pháp đa phần đều dựa trên cơ sở mô hình kiến trúc mạng nơ-ron. Trong tài liệu này, chúng tôi tập trung vào những cải tiến đã có ở hiện tại và khảo sát toàn diện về các phương pháp tóm tắt video dựa trên mạng học sâu. Sau khi trình bày động lực phát triển của công nghệ tóm tắt video, chúng tôi sẽ đưa ra giải thuật và thảo luận về các đặc trưng của giải thuật dựa trên mạng học sâu điển hình. Tiếp theo, chúng tôi sẽ đề xuất phân loại các thuật toán hiện có và phân tích có cơ sở dựa trên những tài liệu liên quan để đưa ra đề xuất cho các phát triển ở trong tương lai. Cuối cùng, chúng tôi đánh giá khách quan các thuật toán tóm tắt video và so sánh hiệu suất của các phương pháp dựa trên mạng học sâu. Từ kết quả so sánh này, cũng như những ý kiến đã được đưa ra trước đó về dữ liệu và tính phù hợp của giao thức đánh giá, chúmg tôi sẽ đưa ra các hướng nghiên cứu tiềm năng của các thuật toán trong tương lai.

Từ khoá—Tóm tắt video, Mạng học sâu nơ-ron, Học có giám sát, Học không có giám sát, Tóm tắt dữ liệu, Giao thức đánh giá

# Giới Thiệu

Youtube được biết đến là một nền tảng chia sẻ video trực tuyến của Mỹ được sử dụng phổ biến rộng rãi trên khắp thế giới. Tháng 7 năm 2015, Youtube công bố rằng họ nhận được hơn 400 giờ nội dung video trên mỗi phút1. Điều này tương đương với 65.7 năm nội dung được tải lên mỗi ngày. Tuy nhiên, Youtube chỉ là một trong các nền tảng chia sẻ video trên thế giới (Tiktok, DailyMotion,…). Ngoài ra còn có các mạng xã hội (Facebook, Instagram, Weibo,…), kho lưu trữ trực tuyến của các tổ chức truyền thông và tin tức đang lưu trữ một khối lượng lớn các nội dung video. Không dừng lại ở đó, các nhà phát triển không ngừng hướng đến nhu cầu của người tiêu dùng hiện tại trên các nền tảng và các thiết bị quay phim trực tuyến (điện thoại thông minh, đồng hồ thông minh,…) bằng cách trang bị cảm biến mạnh mẽ tăng chất lượng video và cho phép tải video đã quay lên Web ngay lập tức. Vì vậy, làm cách nào để người dùng có thể tìm kiếm thông tin một cách nhanh chóng, hiệu quả cao và chính xác giữa vô vàn thông tin như thế này? Câu trả lời cho câu hỏi này không chỉ liên quan đến việc truy xuất thông tin từ video mà còn liên quan đến công nghệ tóm tắt video. Việc tóm tắt giúp cho thông tin được ngắn gọn hơn nhưng mang nội dung trọng tâm của toàn bộ video. Giữa những nội dung video trên Web, việc tóm tắt video hiệu quả sẽ tạo điều kiện thuận lợi cho người phê duyệt và đưa video vào đúng nhóm nội dung của nó. Qua đó tăng lượt tương tác của người xem và mức độ tiêu thị nội dung bằng cách hướng đến đúng đối tượng tiềm năng.

Ứng dụng tóm tắt video được sử dụng trong rất nhiều lĩnh vực. Cụ thể ở các tổ chức truyền thông đa phương tiện, họ sử dụng công nghệ này để chỉ mục, phê duyệt, truy xuất và quảng bá một cách hiệu quả các nội dung video. Từ đó, phân loại video phù hợp với các nhóm đối tượng có nhu cầu. Việc làm này giúp tăng lượt xem cũng như lượng khách hàng tiềm năng tiếp cận sản phẩm để phục vụ cho quảng cáo. Đồng nghĩa với việc tăng lợi nhuận, lượng tiêu thụ và doanh thu cho sản phẩm quảng cáo. Không dừng lại ở lĩnh vực truyền thông, tóm tắt video còn hỗ trợ trong việc tạo ra những bộ phim giới thiệu (trailers) và các tập ngắn của phim truyền hình dài tập. Đặc biệt hơn, tóm tắt video giúp tóm tắt lại nội dung chính đã diễn ra trong ngày hoặc trong tuần của camera giám sát, giúp cho việc giám sát tiến độ và bảo mật hiệu quả hơn về mặt thời gian.

Một số khảo sát về tóm tắt đã xuất hiện trong tài liệu. Một trong những tác phẩm đầu tiên phải kể đến Barbieri và cộng sự (2003) [1] – phân loại các kịch bản của video, đặc điểm của cách tiếp cận với việc tóm tắt video. Trong một nghiên cứu ban đầu khác, Li và cộng sự (2006) [2] đã chia các phương pháp tiếp cận tóm tắt hiện có thành các phương pháp dựa trên tiện ích sử dụng các mô hình chú ý để xác định các đối tượng và cảnh nổi bật, và các phương pháp dựa trên cấu trúc xây dựng dựa trên các cảnh quay và video. Truong và cộng sự (2007) [3] thì lại thảo luận về các thuộc tính ảnh hưởng đến kết quả trong quá trình tóm tắt (miền video, mức độ chi tiết của các đoạn video được sử dụng, phương pháp được sử dụng và loại tóm tắt). Money và cộng sự (2008) [4] phân loại phương pháp thành phương pháp dựa trên phân tích luồng video, phương pháp xử lý siêu dữ liệu video theo ngữ cảnh và cuối cùng là phương pháp kết hợp cả hai. Jiang và cộng sự (2009) [5] thảo luận về một số cách tiếp cận tóm tắt video đặc trưng, bao gồm việc trích xuất các đặc điểm hình ảnh cấp thấp để đánh giá độ tương đồng của frame hình hoặc thực hiện lựa chọn frame hình chính dựa trên phân cụm; phát hiện các sự kiện chính của video bằng cách sử dụng bộ mô tả chuyển động; và xác định cấu trúc video bằng cách sử dụng Eigen-features. Hu và cộng sự (2011) [6] phân loại các phương pháp tóm tắt thành phương pháp nhắm mục tiêu dư thừa trực quan tối thiểu, phương pháp dựa vào phát hiện đối tượng hoặc sự kiện và phương pháp khác dựa trên tích hợp đa phương thức. Ajmal và cộng sự (2012) [7] đã dựa trên các tài liệu liên quan phân loại thành phương pháp dựa trên phân nhóm, phương pháp dựa trên việc phát hiện các sự kiện chính của câu chuyện,… Tuy nhiên, tất cả các tác phẩm nói trên (xuất bản từ năm 2003 đến 2012) đều báo cáo về các cách tiếp cận ban đầu đối với việc tóm tắt video; họ không trình bày bối cảnh tóm tắt đã phát triển như thế nào trong những năm qua và đặc biệt là sau khi giới thiệu các thuật toán học sâu.

Trong các nghiên cứu gần đây hơn của Molino và cộng sự (2017) [8] tập trung chủ yếu vào việc tóm tắt trọng tâm của video và thảo luận về các thông số kỹ thuật cũng như các thách thức. Trong một số công trình nghiên cứu gần đây, Basavarajaiah và cộng sự (2019) [9] đưa ra phân loại các phương pháp tiếp cận tóm tắt khác nhau, bao gồm một số phương pháp dựa trên học sâu được đề cập gần đây. Tuy nhiên, họ chủ yếu tập trung vào các thuật toán tóm tắt có thể áp dụng được trên miền nén. Cuối cùng là cuộc khảo sát của Vivekraj và cộng sự (2019) [10], họ trình bày dựa trên phân loại hai chiều, liên quan đến các phương thức dữ liệu được sử dụng trong quá trình phân tích và kết hợp các khía cạnh của con người. Đối với phương pháp thứ hai, nó tiếp tục chia ra thành phương pháp tạo tóm tắt bằng cách mô hình hóa sự hiểu biết và sở thích của con người (ví dụ là sử dụng các mô hình chú ý, ngữ nghĩa của nội dung trực quan hoặc các chú thích về sự thật cơ bản và các thuật toán học máy), và phương pháp tiếp cận thông thường dựa vào việc xử lý thống kê các tính năng cấp thấp của video. Tuy nhiên, không có cuộc khảo sát nào ở trên trình bày một cách toàn diện những phát triển hiện tại của việc tóm tắt video dựa trên mạng học sâu nơ-ron nâng cao. Trên thực tế, lĩnh vực nghiên cứu này rất hay vì một số phương pháp tiếp cận mới đang được trình bày hằng năm trên các tạp chí và hội nghĩ quốc tế đều được đánh giá cao. Trong cuộc khảo sát này, chúng tôi nghiên cứu chi tiết hơn 40 thuật toán tóm tắt video dựa trên phương pháp học sâu khác nhau được đề xuất trong 5 năm qua. Ngoài ra, so sánh hiệu suất tóm tắt video đã được báo cáo trong các phương pháp tiếp cận dựa trên học sâu gần đây với hiệu suất trong các phương pháp tiếp cận khác [10]-[13] cho thấy rằng trong hầu hết các trường hợp, các phương pháp dựa trên học sâu vượt trội hơn đáng kể so với các phương pháp truyền thống. Qua đó, chúng tôi muốn tiếp tục đóng góp vào công trình phát triển tóm tắt video dựa trên học sâu bằng cách trình bày và thảo kuaajn về các khía cạnh liên quan chẳng hạn như các giao thức được sử dụng để đánh giá.

Bố cục của tài liệu này như sau: Phần II mở đầu bài viết này bằng cách xác định các vấn đề trong tóm tắt video tự động và trình bày các kiểu tóm tắt video nổi bật. Sau đó, chúng tôi sẽ chung cấp mô tả chi tiết hơn về quy trình phân tích của các thuật toán tóm tắt dựa trên học sâu và giới thiệu phân loại theo phương thức dữ liệu được sử dụng, chiến lược đào tạo được chấp nhận và phương pháp tiếp cận học sâu đã được triển khai. Cuối cùng là thảo luận về các khía cạnh liên quan đến phương pháp tóm tắt đã đề cập, chẳng hạn như các thuộc tính mông muốn của bản tóm tắt video tĩnh (dựa trên khung hình) và thời lượng của bản động (dựa trên phân đoạn).

# Các Vấn Đề Liên Quan

**Video summarization** nhằm mục đích tạo bản tóm tắt ngắn nội dung video bằng cách chọn các phần quan trọng và nhiều thông tin nhất của nó. Bản tóm tắt được tạo thường bao gồm một tập hợp các video **key-frames** (khung hình chính) hoặc các **key-fragments** (đoạn video chính) đã được ghép theo thứ tự thời gian để tạo thành một video ngắn hơn.

Loại trước đây của tóm tắt video được gọi là **video storyboard** (bảng phân cảnh video) và loại sau đó được gọi là **video skim** (video lướt). Một lợi thế của video skim so với các bộ frame tĩnh là khả năng bao gồm các yếu tố âm thanh và chuyển động mang lại khả năng tường thuật và biểu cảm, cũng như lượng thông tin được truyền tải cao hơn. Hơn nữa, người xem thường cảm thấy thú vị khi xem một video skim hơn là một slide show (chuỗi trình chiếu) gồm các khung hình [14]. Mặt khác, **storyboard** không bị giới hạn về thời gian hoặc đồng bộ hóa, và do đó, chúng mang lại sự linh hoạt hơn về mặt tổ chức dữ liệu cho các mục đích duyệt và điều hướng [15]-[16].

Sự biểu diễn ở level cao của một quy trình tóm tắt video dựa trên nền tảng deep-learning điển hình được mô tả trong Hình 1.

Diagram

Description automatically generated

*Hình 1. Trình bày cấp cao về quy trình phân tích của các phương pháp tóm tắt video dựa trên deep-learning để tạo video storyboard và video skim.*

Bước đầu tiên của phân tích liên quan đến việc trình bày nội dung trực quan của video với sự trợ giúp của **vectơ đặc trưng**. Thông thường nhất, các vectơ như vậy được trích xuất ở frame-level, cho tất cả các frame hoặc cho một tập hợp con của chúng được chọn thông qua chiến lược lấy mẫu frame (**frame-sampling strategy**) (vd: xử lý 2 khung hình mỗi giây). Bằng cách này, các vectơ đặc trưng được trích xuất lưu trữ thông tin ở mức rất chi tiết và nắm bắt động lực của nội dung trực quan có ý nghĩa cao khi chọn các phần video tạo thành bản tóm tắt.

Thông thường, trong hầu hết các kỹ thuật tóm tắt video dựa trên deep-learning, nội dung trực quan của các khung video được thể hiện bằng các vectơ đặc trưng sâu được trích xuất với sự trợ giúp của mạng nơ-ron được training trước (**pre-trained neural networks**). Vì lý do này, nhiều loại **Convolutional Neural Network** (**CNN**) và **Deep Convolutional Neural Network** (**DCNN**) đã được sử dụng trong danh mục, bao gồm GoogleNet (Inception V1) [17], Inception V3, AlexNet [19], các biến thể của ResNet [20] và các biến thể của VGGnet [21]. Tuy nhiên, **GoogleNet** dường như được sử dụng phổ biến nhất cho đến nay [22]-[51].

Sau đó, các đặc trưng được trích xuất sẽ được sử dụng bởi một mạng tóm tắt sâu (**Deep Summarizer Network**), mạng này được train bằng cách cố gắng giảm thiểu một **objective function** hoặc một tập hợp các objective function.

Đầu ra của **Deep Summarizer Network** sau khi được train có thể là một tập hợp các **key-frame** tạo thành bảng phân cảnh video tĩnh **(static video storyboard**) hoặc một tập hợp các video **key-fragments** được nối với nhau theo thứ tự thời gian và tạo thành một **short video skim**.

Đối với video storyboard, phải tương tự với bộ key-frames do con người lựa chọn và phải thể hiện sự dư thừa hình ảnh tối thiểu. Đối với phần video skim, thường phải bằng hoặc nhỏ hơn độ dài *L* được xác định trước.

Với mục đích thử nghiệm và so sánh, điều này thường được đặt

L = p · T

*trong đó, T là thời lượng video gốc, p là tỷ lệ "phần tóm tắt"/"thời lượng video gốc" (p = 0.15 là giá trị điển hình, nghĩa là phần tóm tắt không quá 15% video gốc)*

Lưu ý, một video skim yêu cầu các segmentation phải liên tục và không chồng chéo, thống nhất về mặt hình ảnh và thời gian ⭢ cung cấp một bản trình bày liền mạch của một phần câu chuyện. Dựa trên segmentation đó và ước tính **importance scores** (điểm số tầm quan trọng) của các frame bởi **Trained Deep Summarizer Network**, điểm số ở video-segment-level được tính bằng cách tính trung bình điểm số của các frame nằm trong mỗi video segment.

Sau đó, các điểm số ở segment-level này được sử dụng để chọn các key-fragments có độ dài tóm tắt *L* và hầu hết các phương pháp giải quyết bước này bằng cách giải quyết vấn đề Knapsack.

Liên quan đến loại dữ liệu được sử dụng, danh sách hiện tại về **video summarization** dựa trên deep-learning có thể được chia thành:

* Các phương pháp tiếp cận đơn phương (**unimodal approache**) chỉ sử dụng phương thức trực quan của video để trích xuất đặc trưng và tìm hiểu tóm tắt theo cách có giám sát (nhưng yếu) hoặc không giám sát.
* Các phương pháp đa phương (**Multimodal method**) khai thác siêu dữ liệu văn bản có sẵn và tìm hiểu tóm tắt theo ngữ nghĩa/danh mục theo cách được giám sát bằng cách tăng mức độ liên quan giữa ngữ nghĩa của bản tóm tắt và ngữ nghĩa của siêu dữ liệu hoặc danh mục video được liên kết

Liên quan đến chiến lược training đã được thông qua, các thuật toán video summarization dựa trên deep-learning có thể được phân loại thô trong các loại sau:

* Các phương pháp tiếp cận được giám sát (**Supervised approache**) dựa trên tập dữ liệu có chú thích sự thật cơ bản do con người gắn nhãn (dưới dạng video tóm tắt, như trong trường hợp của tập dữ liệu SumMe hoặc ở dạng importance scores ở frame-level, như trong trường hợp của tập dữ liệu TVSum [56]), dựa vào đó họ cố gắng khám phá tiêu chí cơ bản cho việc lựa chọn frame/fragment và tóm tắt video.
* Các phương pháp tiếp cận không giám sát (**Unsupervised approache**) khắc phục được nhu cầu về dữ liệu chân thực (mà quá trình sản xuất đòi hỏi thời gian và quy trình chú thích thủ công tốn nhiều công sức), dựa trên cơ chế học hỏi chỉ yêu cầu một bộ sưu tập đủ lớn các video gốc cho quá trình training của họ.
* Các phương pháp tiếp cận bán giám sát (**Weakly-supervised approache**), tương tự như các phương pháp tiếp cận không được giám sát, nhằm mục đích giảm bớt nhu cầu về các bộ dữ liệu được dán nhãn thủ công. Các nhãn yếu ít tốn kém hơn được sử dụng với sự hiểu biết rằng chúng không hoàn hảo so với toàn bộ các chú thích của con người, nhưng vẫn có thể được sử dụng để tạo ra các mô hình dự đoán mạnh mẽ.

Dựa trên các phân loại được mô tả ở trên, phân loại chi tiết hơn của các thư mục tham khảo liên quan được mô tả trong Hình 2. Layer áp chót của minh họa cây thực vật này cho thấy các phương pháp leaning khác nhau đã được áp dụng. Các lá của mỗi node của layer này hiển thị các kỹ thuật được sử dụng để thực hiện từng cách tiếp cận learning và chứa các tham chiếu đến các công trình có liên quan nhất trong thư mục thư khảo. Bảng phân loại này sẽ là cơ sở để trình bày các thư mục tham khảo liên quan trong phần sau.

Diagram

Description automatically generated

# Cách Tiếp Cận Học Sâu

Trong phần này, chúng tôi sẽ giới thiệu ngắn gọn về kiến trúc mạng học sâu (Phần III-A), sau đó tập trung vào ứng dụng của chúng trên lĩnh vực tóm tắt video bằng cách cung cấp đánh giá có hệ thống về các phần liên quan. Đánh giá này bao gồm cách trình bày các lớp khác nhau của phương pháp tóm tắt video có giám sát (Phần III-B), chỉ dựa vào việc phân tích nội dung hình ảnh. Tiếp theo là báo cáo về các phương pháp tiếp cận đa phương thức (Phần III-D) cũng xử lý siêu dữ liệu dựa trên nội dung đã được đưa ra. Cuối cùng là một số nhận xét chung (Phần III-E) tổng quan cách lĩnh vực này phát triển, đặc biệt là trong những năm qua.

## Kiến thức cơ bản về học sâu

Học sâu là một phần của học máy được hình thành bởi sự phát triển mạnh mẽ cũng như đa dạng của dữ liệu và sự tiến bộ vượt bậc trong công nghệ phần cứng. Từ “sâu” ở đây đề cập đến việc sử dụng nhiều lớp trong mạng, qua việc xử lý phi tuyến tính để hiểu được nhiều cấp độ biểu diễn dữ liệu. Việc học này có thể được giám sát hoàn toàn, không giám sát hoặc bán giám sát. Cho đến nay, một số kiến trúc học sâu có thể kể đến: Deep Belief Networks [57], Restricted/Deep Boltzmann Machines [58]-[59], (Variational) Autoencoders [60]-[61], (Deep) Convolutional Neural Networks [62]-[67], … Để có cái nhìn tổng quan hơn về các lớp khác nhau trong kiến trúc mạng học sâu, ta có thể tham khảo các tài liệu sau [68]-[70]. Trong nhiều thập kỷ qua, kiến trúc học sâu đã được sử dụng trong việc xử lý ngôn ngữ tự nhiên [71]-[72], nhận dạng giọng nói [73]-[74], phân tích hình ảnh/video trong y học [75]-[76]. Từ đó, hiệu suất cao và kết quả chính xác cao trong nhiều trường hợp, đạt được khả năng đánh giá tương đương với một chuyên gia con người. Để nắm rõ hơn về chi tiết và ứng dụng, chúng ta có thể xem qua các cuộc khảo sát gần đây [81]-[87].

Tuy nhiên, sự thành công trong thực tế đã tạo nhiều thách thức hơn cho những người nghiên cứu lý thuyết về chủ đề này. Những thách thức có liên quan đến: i) thiết kế kiến trúc học từ các dữ liệu ít được đào tạo, thiếu hoặc bị nhiễu, ii) việc sử dụng các thuật toán tối ưu để điều chỉnh tham số của mạng, iii) việc triển khai các kiến trúc mạng học sâu nhỏ gọn phù hợp cho các thiết bị di động có bộ nhớ thấp, iv) phân tích sự ổn định của mạng học sâu và v) giải thích các cơ chế cơ bản để có thể suy luận như con người trong thời gian thực (lĩnh vực nghiên cứu có thể gọi là AI). Cách để gỉai quyết vấn đề ấy đã được đề cập đến trong một số tác phẩm [88]-[94].

## Tóm tắt video có giám sát

1. **Tìm hiểu tầm quan trọng của frame (khung) bằng cách mô hình hoá sự phụ thuộc theo thời gian giữa các frame.**

Các phương pháp học sâu ban đầu được đưa ra một bài toán dự đoán có có cấu trúc và các ước tính về tầm quan trọng của frame bằng cách mô hình hoá mức độ phụ thuộc theo thời gian của chúng. Như được minh hoạ trong Hình 3, trong giai đoạn đào tạo, Trình tóm tắt nhận đầu vào trình tự của frame video và dữ liệu cơ bản có sẵn, đưa ra tầm quan trọng của từng frame theo sở thích của người dùng. Sau đó, những dữ liệu này được sử dụng để mô hình hoá sự phụ thuộc của các frame hình theo thời gian (được minh hoạ bằng các đường cong cố định) và ước tính tầm quan trọng của các frame hình. Điểm số tầm quan trọng dự đoán được so với dữ liệu cơ sở và kết quả của việc so sánh này hướng đến việc đào tạo Trình tóm tắt. Cách tiếp cận đầu tiên theo hướng này được đề xuất bởi Zhang và cộng sự (2016), sử dụng Long Short-Term Memory (LSTM) để mô hình hoá sự phụ thuộc thời gian trong phạm vi thay đổi giữa các frame hình video. Tầm quan trọng của frame được ước tính bằng cách sử dụng Multi-Layer Perceptron (MLP) và tính đa dạng của nội dung trực quan của bản tóm tắt được tăng lên dựa trên Determinantal Point Process (DPP). Một năm sau, Zhao và cộng sự (2017) đã mô tả kiến trúc LSTM hai lớp. Lớp đầu tiên trích xuất và mã hoá dữ liệu về cấu trúc video. Lớp thứ hai sử dụng thông tin này để ước tính tầm quan trọng ở cấp độ phân đoạn và chọn các phân đoạn chính của video. Trong công trình tiếp theo của họ, Zhao và cộng sự (2018) đã tích hợp một thành phần được đào tạo để xác định cấu trúc thời gian ở cấp độ cảnh quay (shot-level) của video. Kiến thứuc này được áp dụng để ước tính tầm quan trọng ở cấp độ cảnh quay của video dựa trên cảnh quay chính. Trong công trình nghiên cứu cuối cùng của họ, Zhao và cộng sự (2020) đã mở rộng phương pháp bằng cách đưa vào một lớp tensor-train để tránh tính năng ẩn cao từ ma trận ánh xạ. Lớp này được kết hợp với cấu trúc phân cấp của RNN, hoạt động tương tự như lớp trong và ghi nhận sự phụ thuộc theo thời gian của các frame hình nằm trong các ảnh chụp con của video được xác định thủ công (lớp đầu tiên) và trên các ảnh chụp con khác nhau của video (lớp thứu hai). Đầu ra của các lớp này được sử dụng để xác định xác suất của mỗi ảnh con được chọn làm một phần của tóm tắt video. Lebron Casas và cộng sự (2019) đã xây dựng và giới thiệu một cơ chế chú ý để mô hình hóa sự tiến triển theo thời gian của sở thích người dùng. Ở đây, thông tin này được sử dụng để ước tính mức độ quan trọng của frame hình và chọn frame hình chính của video để xây dựng bảng phân cảnh video. Theo hướng tương tự, một số phương pháp đã sử dụng kiến ​​trúc trình tự nối tiếp (được biết đến như seq2seq) kết hợp với các cơ chế chú ý. Ji và cộng sự (2020) tóm tắt video dưới dạng vấn đề học tập seq2seq và đề xuất mạng bộ mã hóa-giải mã dựa trên LSTM với lớp chú ý trung gian. Ji và cộng sự (2020) đã giới thiệu một phần mở rộng của mô hình tóm tắt, từ đó tích hợp một mạng nhúng bảo tồn ngữ nghĩa đánh giá đầu ra của bộ giải mã. Liên quan đến việc bảo toàn ngữ nghĩa của video bằng cách sử dụng một lỗi bảo toàn ngữ nghĩa phù hợp và thay thế cho Mean Square Error đã sử dụng trước đây (MSE) mất mát bởi tổn thất Huber để tăng cường sức mạnh của nó đối với các yếu tố ngoại lai. Sử dụng cơ chế chú ý làm phần cốt lõi của phân tích và nhằm tránh việc sử dụng các LSTM đòi hỏi tính toán cao, Fajtl và cộng sự (2019) đã trình bày một mạng để tóm tắt video, bao gồm một cơ chế tự chú ý mềm và một mạng được kết nối đầy đủ hai lớp để hồi quy điểm số quan trọng của frame hình. Liu và cộng sự (2019) đã mô tả một cách tiếp cận phân cấp kết hợp kiến ​​trúc phân loại (tương tự như trong [32]) như một cơ chế bên trong để ước tính tính đại diện của mỗi cảnh quay và xác định một tập hợp các frame hình chính ứng cử viên. Sau đó, nó sử dụng mô hình chú ý từ nhiều phía để đánh giá thêm tầm quan trọng của ứng viên và chọn các frame chính tạo thành bản tóm tắt. Li và cộng sự (2021) đề xuất một cơ chế chú ý đa dạng toàn cầu bằng cách thực hiện một sự thích nghi của cơ chế tự chú ý của Transformer Network.

1. **Tìm hiểu tầm quan trọng của frame hình bằng cách mô hình hóa cấu trúc không gian của video.**

Nhằm mục đích tìm hiểu cách ước tính tốt hơn về tầm quan trọng của frame hình/phân đoạn video, một số kỹ thuật chú ý đến cả cấu trúc không gian và thời gian của video. Một lần nữa, Trình tóm tắt nhận được đầu vào trình tự của các frame video và dữ liệu cơ bản có sẵn cho biết tầm quan trọng của từng frame theo sở thích của người dùng. Tuy nhiên, mở rộng đường ống phân tích của nhóm phương pháp được mô tả trước đó, nó cũng mô hình hóa sự phụ thuộc không gian giữa các frame (được hiển thị bằng các hình chữ nhật và đường đứt nét trong Hình 3). Một lần nữa, điểm số tầm quan trọng dự đoán được so sánh với dữ liệu trung thực và kết quả của việc so sánh này hướng dẫn việc đào tạo Trình tóm tắt.

Diagram

Description automatically generated

Hình 3. Biểu diễn cấp cao của quy trình phân tích các thuật toán được giám sát thực hiện tóm tắt bằng cách tìm hiểu tầm quan trọng của frame sau khi mô hình hóa sự phụ thuộc theo thời gian hoặc không gian của chúng. Đối với lớp phương pháp thứ hai (tức là mô hình hóa sự phụ thuộc không gian giữa các frame), các hộp giới hạn đối tượng và quan hệ đối tượng theo thời gian được hiển thị bằng các hình chữ nhật và đường đứt nét, được sử dụng để minh họa phần mở rộng mô hình hóa cả sự phụ thuộc theo thời gian và không gian giữa các frame.

Từ quan điểm này, Lal và cộng sự (2019) đã trình bày một kiến ​​trúc bộ mã hóa với các LSTM tích hợp, mô hình hóa mối quan hệ không gian giữa các phần của video. Ngoài các ước tính về tầm quan trọng của frame hình, thuật toán nâng cao tính đa dạng trực quan của bản tóm tắt thông qua cơ chế phát hiện cảnh quay và dự đoán frame hình tiếp theo, dựa trên trực giác rằng các frame hình đầu tiên của một cảnh quay thường có khả năng cao là một phần của bản tóm tắt . Yuan và cộng sự (2019) đã trích xuất các tính năng sâu và nông từ nội dung video bằng cách sử dụng 3D-CNN có thể đào tạo và xây dựng một hình ảnh đại diện mới thông qua chiến lược tổng hợp. Sau đó, họ sử dụng cách biểu diễn này kết hợp với LSTM tích hợp để tạo mô hình cấu trúc không gian và thời gian của video. Cuối cùng, tóm tắt là đã học với sự trợ giúp của một hàm mất mát mới (được gọi là mất mát Sobolev) nhằm mục đích xác định một loạt các điểm số quan trọng cấp frame gần với chuỗi các điểm số trung thực cơ bản bằng cách giảm thiểu khoảng cách của các dẫn xuất của các dữ liệu tuần tự này, và để khai thác cấu trúc thời gian của video. Chu và cộng sự (2019) đã trích xuất thông tin không gian và thời gian bằng cách xử lý các frame thô và bản đồ luồng quang của chúng với CNN, đồng thời học cách ước tính tầm quan trọng của frame dựa trên chú thích của con người và quy trình học phân phối nhãn. Elfeki và cộng sự (2019) kết hợp CNNs và Gated Recurrent Units (một loại RNN) để tạo thành các vectơ đặc điểm không gian, sau đó được sử dụng để ước tính mức độ hoạt động và tầm quan trọng của mỗi frame. Huang và cộng sự (2020) đã đào tạo một mạng nơ-ron để khai thác dữ liệu không gian và sử dụng thông tin trích xuất để tạo ra một đường cong chuyển động giữa các khung. Sau này được sử dụng làm đầu vào cho một phương pháp phát hiện hiệu ứng chuyển tiếp giúp phân đoạn video thành các cảnh quay. Cuối cùng, mô hình tự chú ý khai thác dữ liệu cơ sở do con người tạo ra để tìm hiểu cách ước tính tầm quan trọng trong cảnh quay và chọn các frame/phân đoạn chính của video để tạo thành một bản tóm tắt video tĩnh/động.

1. **Tìm hiểu tóm tắt bằng cách đánh lừa một discriminator (người phân biệt) khi cố gắng phân biệt bản tóm tắt do máy tạo ra với bản tóm tắt do con người tạo ra.**

Theo một cách tiếp cận hoàn toàn khác để giảm thiểu khoảng cách giữa máy được tạo ra và các bản tóm tắt sự thật, một vài phương pháp sử dụng Mạng đối phương chung (GAN). Như đã gửi trước trong Hình 4, Trình tóm tắt (đóng vai trò là Trình tạo GAN) nhận đầu vào trình tự của các frame và tạo ra một bản tóm tắt bằng cách tính toán importance scores ở frame-level.

Diagram

Description automatically generated

Hình 4. Trình bày cấp cao của đường ống phân tích của người được giám sát các thuật toán tìm hiểu tóm tắt với sự trợ giúp của dữ liệu sự thật cơ bản và học đối phương.

Bản tóm tắt được tạo cùng với bản tóm tắt video tối ưu cho video này được cung cấp làm đầu vào cho **discriminator** có thể đào tạo để tạo ra điểm số định lượng sự giống nhau của chúng. Việc đào tạo toàn bộ kiến ​​trúc tóm tắt được thực hiện theo phương thức đối nghịch. Trình tóm tắt cố gắng đánh lừa **discriminator** để không phân biệt được dự đoán với tóm tắt do người dùng tạo và **discriminator** hướng đến việc tìm hiểu cách tạo ra sự phân biệt này. Khi độ tin cậy của **discriminator** đối xử rất thấp (nghĩa là lỗi phân loại gần bằng nhau cho bản tóm tắt do máy và người dùng tạo), thì người tóm tắt có thể tạo một bản tóm tắt rất gần với mong đợi của người dùng. Trong bối cảnh này, Zhang et al. (2019) đã đề xuất một phương pháp kết hợp LSTM và đơn vị quan hệ thời gian pha loãng (DTR) để ước tính mức độ phụ thuộc theo thời gian giữa các khung ở các cửa sổ thời gian khác nhau và tìm hiểu tóm tắt bằng cách cố gắng đánh lừa một trình phân loại có thể đào tạo khi phân biệt tóm tắt dựa trên máy từ sự thật cơ bản và một sự thật được tạo ra một cách ngẫu nhiên. Trong một tác phẩm khác cùng năm, Fu et al. (2019) đã đề xuất một phương pháp học tập đối nghịch cho bản tóm tắt video có giám sát (bán phần). Trình tạo / Trình tóm tắt (Generator/Summarizer) là Mạng con trỏ dựa trên sự chú ý xác định điểm bắt đầu và điểm kết thúc của mỗi đoạn video được sử dụng để tạo thành phần tóm tắt. Discriminator là một công cụ phân loại 3D-CNN để đánh giá xem một phân đoạn là từ sự thật cơ bản hay bản tóm tắt do máy tạo ra. Thay vì sử dụng phần thua đối đầu thông thường, trong thuật toán này, đầu ra của Bộ phân biệt được sử dụng như một phần thưởng để đào tạo Bộ tạo/Bộ tổng hợp dựa trên việc học tăng cường. Cho đến nay, việc sử dụng GAN để tóm tắt video được giám sát còn hạn chế. Tuy nhiên, framework machine learning này đã được sử dụng rộng rãi để tóm tắt video không có giám sát, như được thảo luận trong phần sau.

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Đánh Giá Tóm Tắt Video

## Tập dữ liệu

Bốn tập dữ liệu (**dataset**) chiếm ưu thế trong tham khảo video summarization: SumMe, TVSum, OVP và Youtube.

* SumMe bao gồm 25 video có thời lượng từ 1 đến 6 phút, với nội dung video đa dạng, được quay từ cả hai góc nhìn, góc thứ nhất và góc thứ ba. Mỗi video đã được chú thích bởi 15 đến 18 người dùng dưới dạng các **key-fragments** và do đó được liên kết với nhiều bản tóm tắt của người dùng ở cấp độ phân đoạn (**fragment-level**) có độ dài từ 5% đến 15% thời lượng video ban đầu.
* TVSum bao gồm 50 video có thời lượng từ 1 đến 11 phút, chứa nội dung video từ 10 danh mục của dataset TRECVID MED. Các video TVSum đã được 20 người dùng chú thích dưới dạng **importance scores** ở cấp độ khung hình (dao động từ 1 đến 5).
* Cả OVP và Youtube đều chứa 50 video, có chú thích là tập hợp các **key-frame**, do 5 người dùng tạo ra. Thời lượng video từ 1 đến 4 phút đối với OVP và từ 1 đến 10 phút đối với Youtube. Cả hai dataset đều bao gồm các video có nội dung video đa dạng, chẳng hạn như phim tài liệu, video giáo dục, phù du, lịch sử và bài giảng (OVP) và phim hoạt hình, tin tức, thể thao, quảng cáo, TV-shows và video gia đình (Youtube).

Với kích thước của từng dataset này, chúng tôi thảo luận rằng thiếu các dataset được chú thích quy mô lớn có thể hữu ích để cải thiện việc đào tạo các kiến trúc deep learning có giám sát phức tạp.

Một số dataset ít được sử dụng hơn để tóm tắt video là CoSum, MED-Summary, Video Titles in the Wild (VTW), League of Legends (LoL) và FVPSum.

* CoSum đã được tạo để đánh giá tóm tắt đồng video. Nó bao gồm 51 video được tải xuống từ YouTube bằng cách sử dụng 10 thuật ngữ truy vấn liên quan đến nội dung video của dataset SumMe. Mỗi video dài khoảng 4 phút và được chú thích bởi 3 người dùng khác nhau, những người đã chọn tập hợp các **key-fragments**.
* MED chứa 160 video được chú thích từ dataset TRECVID 2011 MED. 60 video tạo thành tập hợp xác thực (từ 15 danh mục sự kiện) và 100 video còn lại tạo thành tập kiểm tra (từ 10 danh mục sự kiện), với hầu hết thời lượng từ 1 đến 5 phút. Các chú thích là một tập hợp các **importance scores**, được tính trung bình trên 1 đến 4 chú thích.
* Theo như tập dữ liệu VTW có liên quan, nó bao gồm 18100 video miền mở, với 2000 video trong số đó được chú thích dưới dạng điểm đánh dấu **(highlight score)** ở subshot-level. Các video này là do người dùng tạo, không có thời lượng, chứa sự kiện nổi bật và có thời lượng trung bình là 1,5 phút.
* Liên quan đến LoL, nó có 218 video dài (30 đến 50 phút), hiển thị các trận đấu của trò chơi từ Giải vô địch Liên minh huyền thoại Bắc Mỹ (NALCS). Các chú thích bắt nguồn từ một kênh Youtube cung cấp các video nổi bật do cộng đồng tạo (video có thời lượng từ 5 đến 7 phút). Do đó, một tập hợp các key-fragment có sẵn cho mỗi video.
* Cuối cùng, FPVSum là datasey tóm tắt video ở góc nhìn thứ nhất. Nó chứa 98 video (tổng thời lượng hơn 7 giờ) từ 14 danh mục video thân thiện với người xem của GoPro. Đối với mỗi danh mục, khoảng 35% chuỗi video được ít nhất 10 người dùng chú thích với điểm trung thực cơ bản (**ground-truth scores**), trong khi phần còn lại được xem như các ví dụ không được gắn nhãn. Các đặc điểm chính của từng dataset đã thảo luận ở trên, được trình bày ngắn gọn trong Bảng I.

Điều đáng nói là trong công việc này, chúng tôi chỉ tập trung vào các dataset phù hợp để đánh giá các phương pháp tóm tắt video, cụ thể là các dataset có chứa các chú thích về nội dung cơ bản liên quan đến tóm tắt hoặc tầm quan trọng ở frame/fragment-level của mỗi video. **Bảng II** tóm tắt các dataset được sử dụng bởi các phương pháp tiếp cận dựa trên deep learning để tổng hợp video. Từ bảng này, rõ ràng cho đến nay, SumMe và TVSum là những dataset được sử dụng phổ biến nhất. OVP và Youtube cũng được sử dụng trong một số hoạt động, nhưng chủ yếu cho mục đích tăng dữ liệu.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

1. M. Barbieri, L. Agnihotri, and N. Dimitrova, “Video summarization: methods and landscape,” in Internet Multimedia Management Systems IV, J. R. Smith, S. Panchanathan, and T. Zhang, Eds., vol. 5242, International Society for Optics and Photonics. SPIE, 2003, pp. 1 – 13.
2. Ying Li, Shih-Hung Lee, Chia-Hung Yeh, and C.-J. Kuo, “Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques,” IEEE Signal Processing Magazine, vol. 23, no. 2, pp. 79–89, 2006.
3. B. T. Truong and S. Venkatesh, “Video Abstraction: A Systematic Review and Classification,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 3, no. 1, p. 3–es, Feb. 2007.
4. A. G. Money and H. Agius, “Video summarisation: A conceptual framework and survey of the state of the art,” Journal of Visual Communication and Image Representation, vol. 19, no. 2, pp. 121 – 143, 2008.
5. R. M. Jiang, A. H. Sadka, and D. Crookes, Advances in Video Summarization and Skimming. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 27–50.
6. W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, “A Survey on Visual Content-Based Video Indexing and Retrieval,” IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 41, no. 6, pp. 797–819, 2011.
7. M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas, and F. A. Shah, “Video Summarization: Techniques and Classification,” in Computer Vision and Graphics, L. Bolc, R. Tadeusiewicz, L. J. Chmielewski, and K. Wojciechowski, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–13.
8. A. G. del Molino, C. Tan, J. Lim, and A. Tan, “Summarization of Egocentric Videos: A Comprehensive Survey,” IEEE Trans. on HumanMachine Systems, vol. 47, no. 1, pp. 65–76, Feb 2017.
9. M. Basavarajaiah and P. Sharma, “Survey of Compressed Domain Video Summarization Techniques,” ACM Computing Surveys, vol. 52, no. 6, Oct. 2019.
10. V. V. K., D. Sen, and B. Raman, “Video Skimming: Taxonomy and Comprehensive Survey,” ACM Computing Surveys, vol. 52, no. 5, Sep. 2019.
11. M. Gygli, H. Grabner, and L. V. Gool, “Video summarization by learning submodular mixtures of objectives,” in 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 3090–3098.
12. K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, “Summary Transfer: Exemplar-Based Subset Selection for Video Summarization,” in 2016 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1059–1067.
13. M. Ma, S. Mei, S. Wan, Z. Wang, D. D. Feng, and M. Bennamoun, “Similarity based block sparse subset selection for video summarization,” IEEE Trans. on Circuits and Systems for Video Technology, pp. 1–1, 2020.
14. Y. Li, T. Zhang, and D. Tretter, “An overview of video abstraction techniques,” Hewlett Packard, Technical Reports, 01 2001.
15. J. Calic, D. P. Gibson, and N. W. Campbell, “Efficient Layout of Comic-Like Video Summaries,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 17, no. 7, pp. 931–936, July 2007.
16. T. Wang, T. Mei, X. Hua, X. Liu, and H. Zhou, “Video Collage: A Novel Presentation of Video Sequence,” in 2007 IEEE Int. Conf. on Multimedia and Expo, July 2007, pp. 1479–1482.
17. C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 1–9.
18. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in 2016 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.
19. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
20. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
21. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Int. Conf. on Learning Representations, 2015
22. K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, “Video Summarization with Long Short-Term Memory,” in Europ. Conf. on Computer Vision (ECCV) 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 766–782
23. B. Zhao, X. Li, and X. Lu, “Hierarchical Recurrent Neural Network for Video Summarization,” in Proc. of the 2017 ACM on Multimedia Conf. (MM ’17). New York, NY, USA: ACM, 2017, pp. 863–871.
24. L. Lebron Casas and E. Koblents, “Video Summarization with LSTM and Deep Attention Models,” in MultiMedia Modeling, I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W.-H. Cheng, and S. Vrochidis, Eds. Cham: Springer International Publishing, 2019, pp. 67–79.
25. J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, “Summarizing Videos with Attention,” in Asian Conf. on Computer Vision (ACCV) 2018 Workshops, G. Carneiro and S. You, Eds. Cham: Springer International Publishing, 2019, pp. 39–54.
26. Z. Ji, K. Xiong, Y. Pang, and X. Li, “Video Summarization With Attention-Based Encoder–Decoder Networks,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 30, no. 6, pp. 1709–1717, 2020.
27. Z. Ji, F. Jiao, Y. Pang, and L. Shao, “Deep attentive and semantic preserving video summarization,” Neurocomputing, vol. 405, pp. 200 – 207, 2020.
28. M. Rochan, L. Ye, and Y. Wang, “Video Summarization Using Fully Convolutional Sequence Networks,” in Europ. Conf. on Computer Vision (ECCV) 2018, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 358–374.
29. L. Feng, Z. Li, Z. Kuang, and W. Zhang, “Extractive Video Summarizer with Memory Augmented Neural Networks,” in Proc. of the 26th ACM Int. Conf. on Multimedia (MM ’18). New York, NY, USA: ACM, 2018, pp. 976–983.
30. J. Wang, W. Wang, Z. Wang, L. Wang, D. Feng, and T. Tan, “Stacked Memory Network for Video Summarization,” in Proc. of the 27th ACM Int. Conf. on Multimedia (MM ’19). New York, NY, USA: ACM, 2019, p. 836–844.
31. M. Elfeki and A. Borji, “Video Summarization Via Actionness Ranking,” in IEEE Winter Conf. on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, January 7-11, 2019, Jan 2019, pp. 754–763.
32. B. Mahasseni, M. Lam, and S. Todorovic, “Unsupervised Video Summarization with Adversarial LSTM Networks,” in 2017 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2982–2991.
33. E. Apostolidis, A. I. Metsai, E. Adamantidou, V. Mezaris, and I. Patras, “A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization,” in Proc. of the 1st Int. Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV ’19). New York, NY, USA: ACM, 2019, pp. 17–25.
34. T. Fu, S. Tai, and H. Chen, “Attentive and Adversarial Learning for Video Summarization,” in IEEE Winter Conf. on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, January 7-11, 2019, 2019, pp. 1579–1587.
35. Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon, “Discriminative feature learning for unsupervised video summarization,” in Proc. of the 2019 AAAI Conf. on Artificial Intelligence, 2019.
36. L. Yuan, F. E. H. Tay, P. Li, L. Zhou, and J. Feng, “Cycle-SUM: CycleConsistent Adversarial LSTM Networks for Unsupervised Video Summarization,” in Proc. of the 2019 AAAI Conf. on Artificial Intelligence, 2019.
37. E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, “Unsupervised Video Summarization via Attention-Driven Adversarial Learning,” in Proc. of the 26th Int. Conf. on Multimedia Modeling (MMM 2020). Cham: Springer International Publishing, 2020, pp. 492–504.
38. X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan, “Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks,” in Proc. of the 27th ACM Int. Conf. on Multimedia (MM ’19). New York, NY, USA: ACM, 2019, pp. 2296–2304.
39. M. Rochan and Y. Wang, “Video Summarization by Learning From Unpaired Data,” in 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2019, pp. 7894–7903.
40. K. Zhou and Y. Qiao, “Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward,” in Proc. of the 2018 AAAI Conf. on Artificial Intelligence, 2018.
41. N. Gonuguntla, B. Mandal, N. Puhan et al., “Enhanced Deep Video Summarization Network,” in 2019 British Machine Vision Conf. (BMVC), 2019.
42. B. Zhao, X. Li, and X. Lu, “Property-constrained dual learning for video summarization,” IEEE Trans. on Neural Networks and Learning Systems, vol. 31, no. 10, pp. 3989–4000, 2020.
43. S. Cai, W. Zuo, L. S. Davis, and L. Zhang, “Weakly-Supervised Video Summarization Using Variational Encoder-Decoder and Web Prior,” in Europ. Conf. on Computer Vision (ECCV) 2018, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 193–210.
44. Y. Chen, L. Tao, X. Wang, and T. Yamasaki, “Weakly Supervised Video Summarization by Hierarchical Reinforcement Learning,” in Proc. of the ACM Multimedia Asia (MMAsia ’19). New York, NY, USA: ACM, 2019.
45. K. Zhou, T. Xiang, and A. Cavallaro, “Video Summarisation by Classification with Deep Reinforcement Learning,” in 2018 British Machine Vision Conf. (BMVC), 2018.
46. Y. Yuan, T. Mei, P. Cui, and W. Zhu, “Video Summarization by Learning Deep Side Semantic Embedding,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 29, no. 1, pp. 226–237, Jan 2019.
47. E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, “AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization,” IEEE Trans. on Circuits and Systems for Video Technology, pp. 1–1, 2020.
48. Y. Jung, D. Cho, S. Woo, and I. S. Kweon, “Global-and-local relative position embedding for unsupervised video summarization,” in Europ. Conf. on Computer Vision (ECCV) 2020, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 167–183.
49. G. Yaliniz and N. Ikizler-Cinbis, “Using independently recurrent networks for reinforcement learning based unsupervised video summarization,” Multimedia Tools and Applications, vol. 80, no. 12, pp. 17 827–17 847, 2021. [Online]. Available: <https://doi.org/10.1007/s11042-020-10293-x>
50. P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao, “Exploring global diverse attention via pairwise temporal relation for video summarization,” Pattern Recognition, vol. 111, p. 107677, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320320304805>
51. B. Zhao, X. Li, and X. Lu, “TTH-RNN: Tensor-Train Hierarchical Recurrent Neural Network for Video Summarization,” IEEE Trans. on Industrial Electronics, vol. 68, no. 4, pp. 3629–3637, 2020.
52. S. Lal, S. Duggal, and I. Sreedevi, “Online video summarization: Predicting future to better summarize present,” in 2019 IEEE Winter Conf. on Applications of Computer Vision (WACV). IEEE, 2019, pp. 471–480.
53. W.-T. Chu and Y.-H. Liu, “Spatiotemporal Modeling and Label Distribution Learning for Video Summarization,” in 2019 IEEE 21st Int. Workshop on Multimedia Signal Processing (MMSP). IEEE, 2019, pp. 1–6.
54. Y. Zhang, M. Kampffmeyer, X. Zhao, and M. Tan, “DTR-GAN: Dilated Temporal Relational Adversarial Network for Video Summarization,” in Proc. of the ACM Turing Celebration Conf. (ACM TURC ’19) - China. New York, NY, USA: ACM, 2019, pp. 89:1–89:6.
55. M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, “Creating Summaries from User Videos,” in Europ. Conf. on Computer Vision (ECCV) 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 505–520. [Online]. Available: <https://gyglim.github.io/me/>
56. Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, “TVSum: Summarizing web videos using titles,” in 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 5179–5187. [Online]. Available: <https://github.com/yalesong/tvsum>
57. G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, p. 1527–1554, Jul. 2006. [Online]. Available: <https://doi.org/10.1162/neco.2006.18.7.1527>
58. R. Salakhutdinov, A. Mnih, and G. Hinton, “Restricted boltzmann machines for collaborative filtering,” in Proc. of the 24th Int. Conf. on Machine Learning, ser. ICML ’07. New York, NY, USA: Association for Computing Machinery, 2007, p. 791–798. [Online]. Available: <https://doi.org/10.1145/1273496.1273596>
59. R. Salakhutdinov and G. Hinton, “Deep boltzmann machines,” in Proc. of the 12th Int. Conf. on Artificial Intelligence and Statistics, ser. Proc. of Machine Learning Research, D. van Dyk and M. Welling, Eds., vol. 5. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR, 16–18 Apr 2009, pp. 448–455. [Online]. Available: <http://proceedings.mlr.press/v5/salakhutdinov09a.html>
60. G. E. Hinton and R. S. Zemel, “Autoencoders, minimum description length and helmholtz free energy,” in Proc. of the 6th Int. Conf. on Neural Information Processing Systems, ser. NIPS’93. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, p. 3–10.
61. D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” in 2nd Int. Conf. on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.
62. Y. LeCun and Y. Bengio, Convolutional Networks for Images, Speech, and Time Series. Cambridge, MA, USA: MIT Press, 1998, p. 255–258.
63. C. Goller and A. Kuchler, “Learning task-dependent distributed representations by backpropagation through structure,” in Proc. of the 1996 Int. Conf. on Neural Networks (ICNN’96), vol. 1, 1996, pp. 347–352 vol.1.
64. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning Representations by Back-Propagating Errors. Cambridge, MA, USA: MIT Press, 1988, p. 696–699.
65. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. of the 27th Int. Conf. on Neural Information Processing Systems - Volume 2, ser. NIPS’14. Cambridge, MA, USA: MIT Press, 2014, p. 2672–2680.
66. F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” IEEE Trans. on Neural Networks, vol. 20, no. 1, pp. 61–80, 2009.
67. J. Gast and S. Roth, “Lightweight probabilistic deep networks,” in 2018 IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. IEEE Computer Society, 2018, pp. 3369–3378. [Online]. Available: http://openaccess.thecvf.com/content cvpr 2018/html/Gast Lightweight Probabilistic
68. W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, “A survey of deep neural network architectures and their applications,” Neurocomputing, vol. 234, pp. 11–26, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231216315533>
69. S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. S. Iyengar, “A survey on deep learning: Algorithms, techniques, and applications,” ACM Computing Surveys, vol. 51, no. 5, Sep. 2018. [Online]. Available: <https://doi.org/10.1145/3234150>
70. S. Dong, P. Wang, and K. Abbas, “A survey on deep learning and its applications,” Computer Science Review, vol. 40, p. 100379, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1574013721000198>
71. H. Schwenk, “Continuous space translation models for phrasebased statistical machine translation,” in Proc. of COLING 2012: Posters. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 1071–1080. [Online]. Available: <https://www.aclweb.org/anthology/C12-2104>
72. L. Dong, F. Wei, K. Xu, S. Liu, and M. Zhou, “Adaptive multi-compositionality for recursive neural network models,” IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 24, no. 3, p. 422–431, Mar. 2016. [Online]. Available: <https://doi.org/10.1109/TASLP.2015.2509257>
73. Y. You, Y. Qian, T. He, and K. Yu, “An investigation on dnn-derived bottleneck features for gmm-hmm based robust speech recognition,” in 2015 IEEE China Summit and Int. Conf. on Signal and Information Processing (ChinaSIP), 2015, pp. 30–34.
74. A. L. Maas, P. Qi, Z. Xie, A. Y. Hannun, C. T. Lengerich, D. Jurafsky, and A. Y. Ng, “Building dnn acoustic models for large vocabulary speech recognition,” Comput. Speech Lang., vol. 41, no. C, p. 195–213, Jan. 2017. [Online]. Available: <https://doi.org/10.1016/j.csl.2016.06.007>
75. K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, “Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images,” IEEE Trans. on Medical Imaging, vol. 35, no. 5, pp. 1196–1206, 2016.
76. T. Liu, Q. Meng, A. Vlontzos, J. Tan, D. Rueckert, and B. Kainz, “Ultrasound video summarization using deep reinforcement learning,” in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham: Springer International Publishing, 2020, pp. 483–492.
77. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
78. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1409.1556>
79. H. Xiao, J. Feng, G. Lin, Y. Liu, and M. Zhang, “Monet: Deep motion exploitation for video object segmentation,” in 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1140– 1148.
80. P. Xu, M. Ye, X. Li, Q. Liu, Y. Yang, and J. Ding, “Dynamic background learning through deep auto-encoder networks,” in Proc. of the 22nd ACM Int. Conf. on Multimedia, ser. MM ’14. New York, NY, USA: Association for Computing Machinery, 2014, p. 107–116. [Online]. Available: <https://doi.org/10.1145/2647868.2654914>
81. J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, “Recent advances in convolutional neural networks,” Pattern Recognition, vol. 77, pp. 354–377, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320317304120>
82. T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, “Deep neural network concepts for background subtraction:a systematic review and comparative evaluation,” Neural Networks, vol. 117, pp. 8–66, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0893608019301303>
83. P. Dixit and S. Silakari, “Deep learning algorithms for cybersecurity applications: A technological and status review,” Computer Science Review, vol. 39, p. 100317, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1574013720304172>
84. L. Zhang, M. Wang, M. Liu, and D. Zhang, “A survey on deep learning for neuroimaging-based brain disorder analysis,” Frontiers in Neuroscience, vol. 14, p. 779, 2020. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fnins.2020.00779>
85. J. Gao, P. Li, Z. Chen, and J. Zhang, “A survey on deep learning for multimodal data fusion,” Neural Computation, vol. 32, no. 5, pp. 829–864, 2020.
86. J. Li, A. Sun, J. Han, and C. Li, “A survey on deep learning for named entity recognition,” IEEE Trans. on Knowledge and Data Engineering, pp. 1–1, 2020.
87. S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, “A survey of deep learning techniques for autonomous driving,” Journal of Field Robotics, vol. 37, no. 3, pp. 362–386, 2020. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21918>
88. K. K. Thekumparampil, A. Khetan, Z. Lin, and S. Oh, “Robustness of conditional gans to noisy labels,” in Proc. of the 32nd Int. Conf. on Neural Information Processing Systems, ser. NIPS’18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 10292–10303.
89. A. Creswell and A. A. Bharath, “Denoising adversarial autoencoders,” IEEE Trans. on Neural Networks and Learning Systems, vol. 30, no. 4, pp. 968–984, 2019.
90. N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, “Learning with noisy labels,” in Advances in Neural Information Processing Systems, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013. [Online]. Available: <https://proceedings.neurips.cc/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Pa>
91. G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” 2018.
92. M. Aubry and B. C. Russell, “Understanding deep features with computer-generated imagery,” in 2015 IEEE Int. Conf. on Computer Vision (ICCV), 2015, pp. 2875–2883.
93. C. Yun, S. Sra, and A. Jadbabaie, “A critical view of global optimality in deep learning,” in Int. Conf. on Machine Learning Representations, 2018.
94. Z. Zheng and P. Hong, “Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks,” in Proceedings of the 32nd International Conference on Neural Information Processing Systems, ser. NIPS’18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 7924–7933.
95. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
96. A. Kulesza and B. Taskar, Determinantal Point Processes for Machine Learning. Hanover, MA, USA: Now Publishers Inc., 2012.
97. B. Zhao, X. Li, and X. Lu, “HSA-RNN: Hierarchical StructureAdaptive RNN for Video Summarization,” in 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7405– 7414.
98. Y.-T. Liu, Y.-J. Li, F.-E. Yang, S.-F. Chen, and Y.-C. F. Wang, “Learning Hierarchical Self-Attention for Video Summarization,” in 2019 IEEE Int. Conf. on Image Processing (ICIP). IEEE, 2019, pp. 3377–3381.
99. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, “Attention is All You Need,” in Proc. of the 31st Int. Conf. on Neural Information Processing Systems (NIPS’17). Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.
100. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.
101. L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834–848, 2018.
102. Y. Yuan, H. Li, and Q. Wang, “Spatiotemporal modeling for video summarization using convolutional recurrent neural network,” IEEE Access, vol. 7, pp. 64 676–64 685, 2019.
103. K. Cho, B. van Merri¨enboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” in Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734.
104. C. Huang and H. Wang, “A Novel Key-Frames Selection Framework for Comprehensive Video Summarization,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 577–589, 2020.
105. O. Vinyals, M. Fortunato, and N. Jaitly, “Pointer Networks,” in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2692–2700.
106. S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, “Independently recurrent neural network (indrnn): Building a longer and deeper rnn,” 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5457–5466, 2018.
107. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal Segment Networks: Towards Good Practices for Deep Action Recognition,” in Europ. Conf. on Computer Vision (ECCV) 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 20–36.
108. Y. Zhang, X. Liang, D. Zhang, M. Tan, and E. P. Xing, “Unsupervised object-level video summarization with online motion auto-encoder,” Pattern Recognition Letters, 2018.
109. R. Panda, A. Das, Z. Wu, J. Ernst, and A. K. Roy-Chowdhury, “Weakly Supervised Summarization of Web Videos,” in 2017 IEEE Int. Conf. on Computer Vision (ICCV), Oct 2017, pp. 3677–3686.