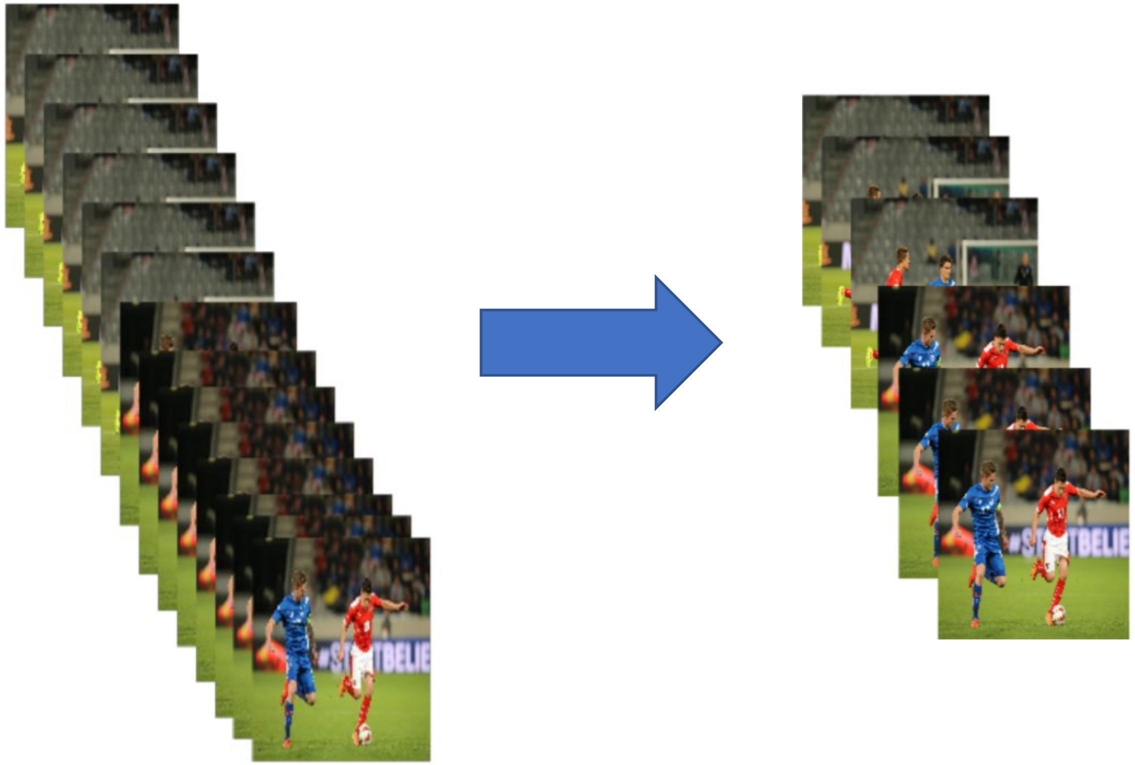
Video Summarization Using Deep Neural Networks

Speaker: Anh Tran

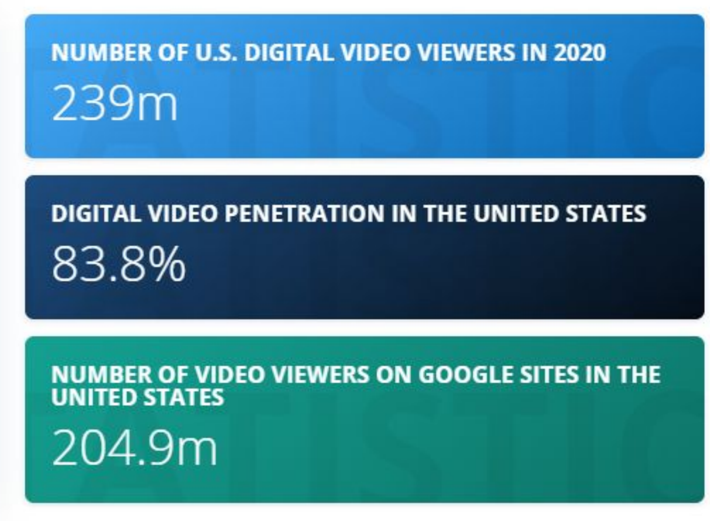
Background Information

Video Summarization

- Input: raw frames in a long video
- Output: subset of selected frames (shots) as a representative summary of video content



Why Video Summarization?



NUMBER OF U.S. DIGITAL VIDEO VIEWERS IN 2020

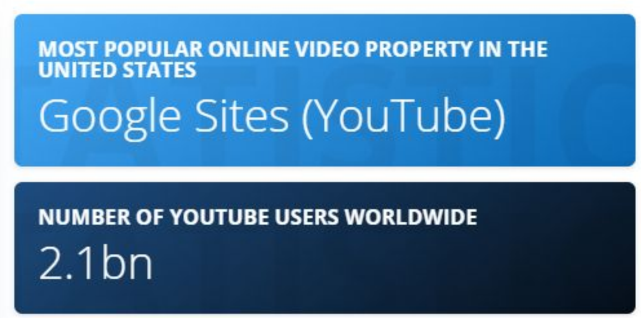
239m

DIGITAL VIDEO PENETRATION IN THE UNITED STATES

83.8%

NUMBER OF VIDEO VIEWERS ON GOOGLE SITES IN THE UNITED STATES

204.9m

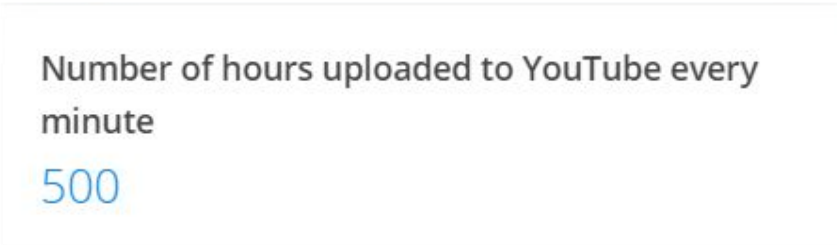


MOST POPULAR ONLINE VIDEO PROPERTY IN THE UNITED STATES

Google Sites (YouTube)

NUMBER OF YOUTUBE USERS WORLDWIDE

2.1bn



Number of hours uploaded to YouTube every minute

500

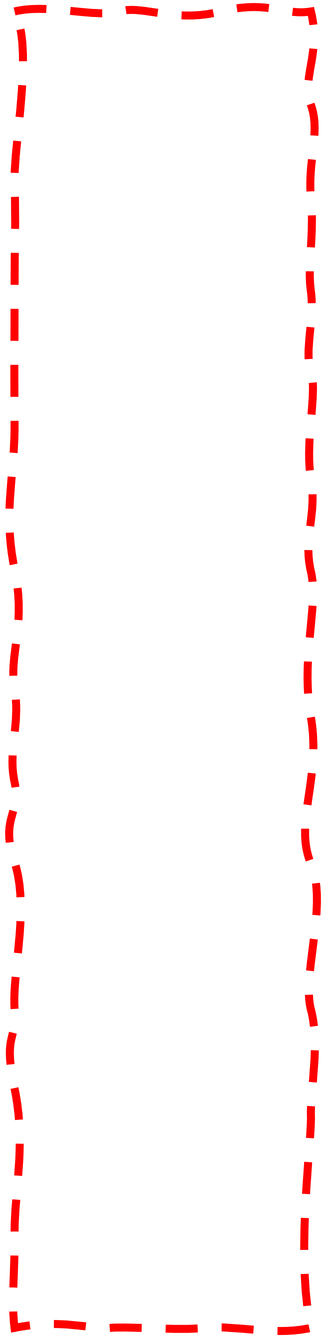
Applications of Video Summarization

- For media organizations: allow for effective indexing, browsing, retrieval and promotion of entertainment media assets
- For video sharing platforms: improve viewing experience, enhance viewers’ engagement and increase content consumption.
- Generate trailers or teasers of movies or TV series
- Generate video synopsis of surveillance camera, for time-efficient progress monitoring or security purposes
- Generate highlights of event (sports game, performance, public debate, etc.)

Video Summarization
Using Deep Neural Networks

Overview

Overview:
Visual content as Feature Vectors



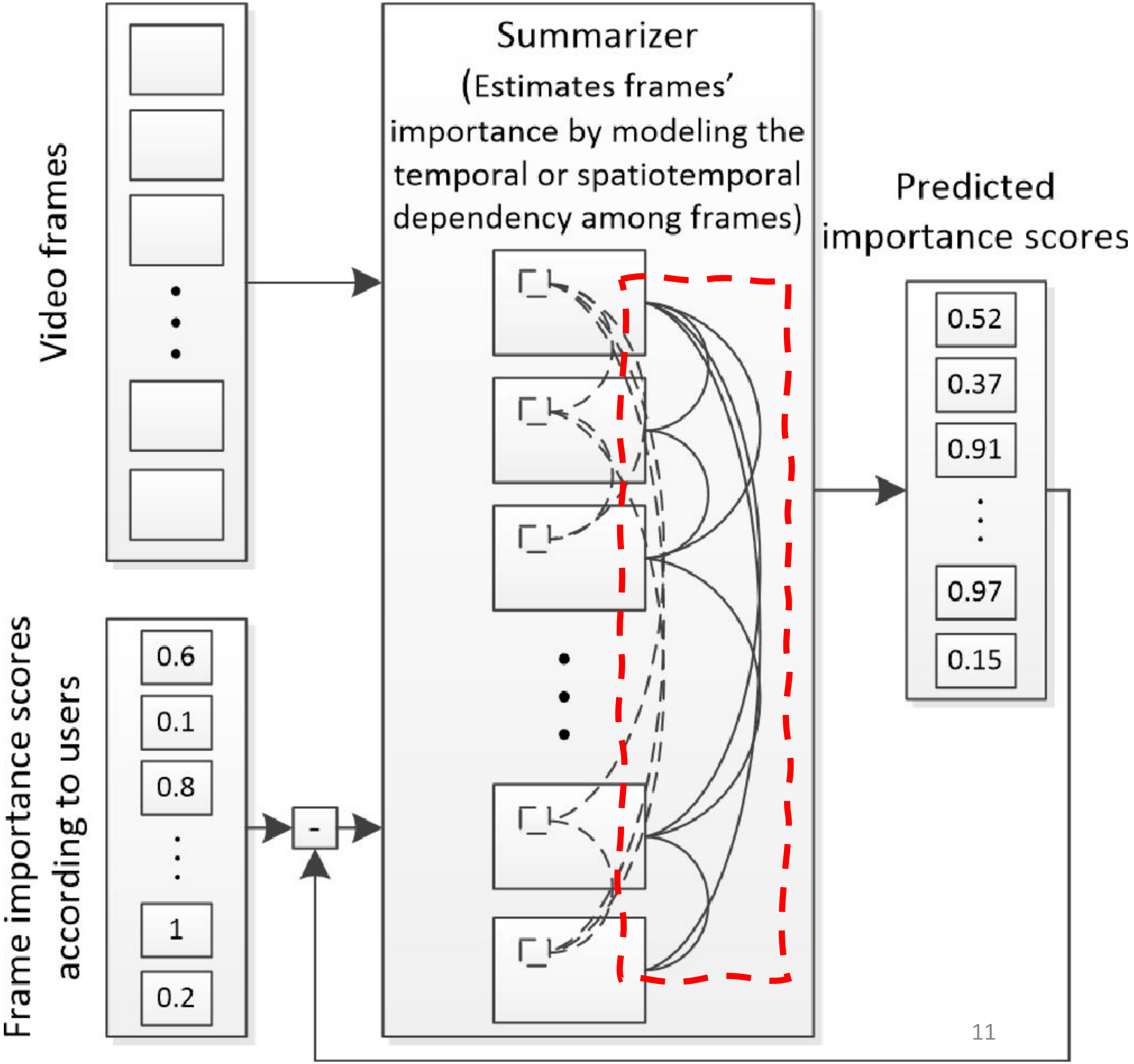
Overview: Deep Summarizer Network

Overview: Approaches

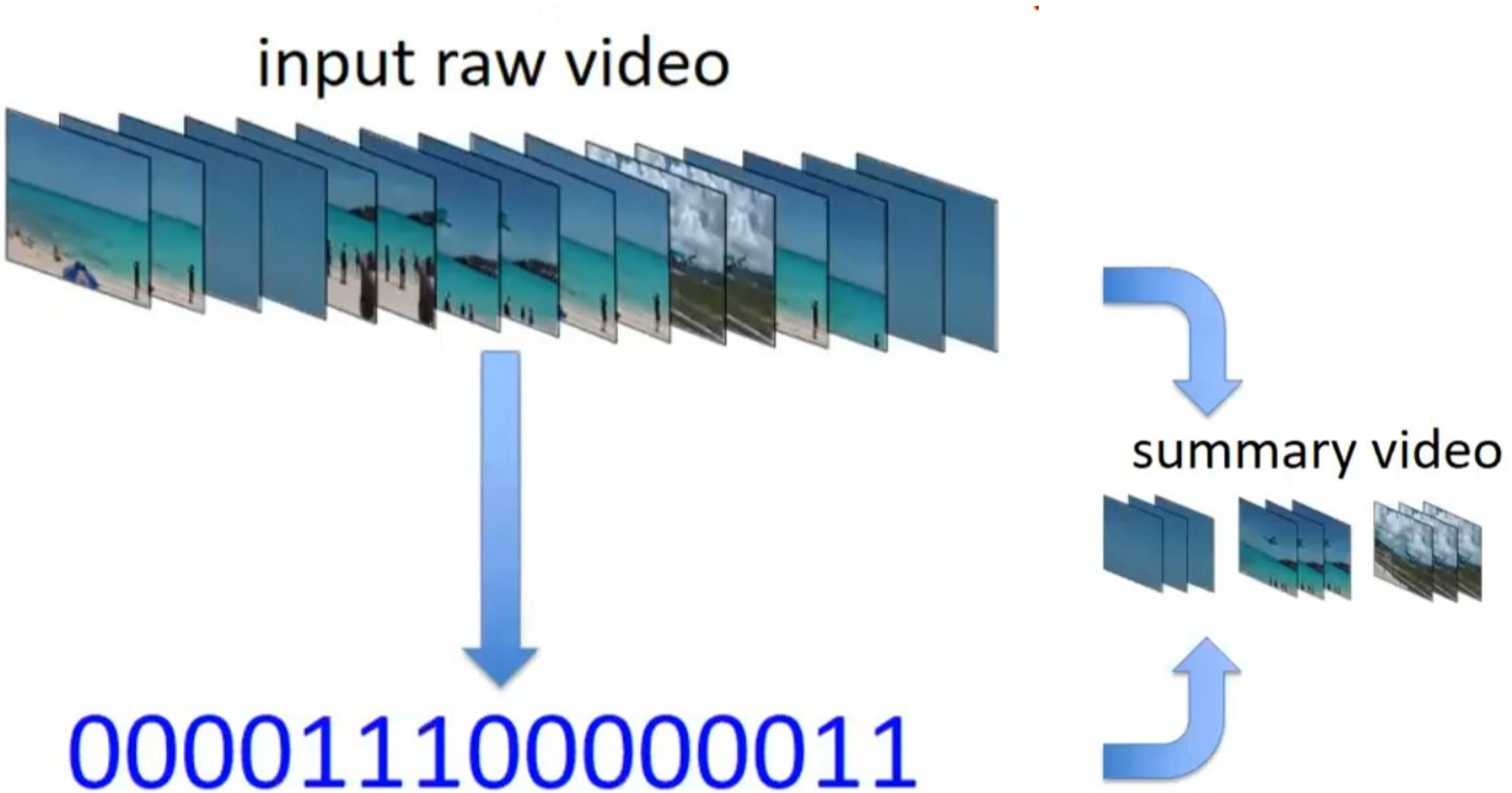
Unimodal approaches

Supervised Learning

- Learn frame importance by modeling the **temporal** dependency among frames

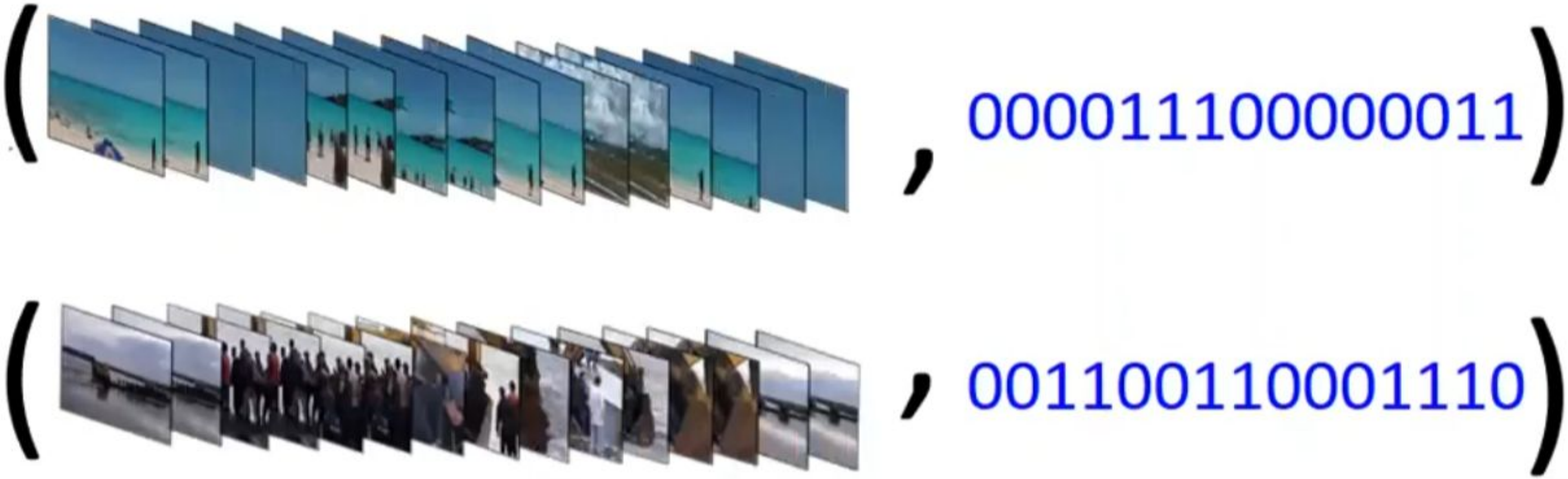

Video frames
Summarizer
(Estimates frames' importance by modeling the temporal or spatiotemporal dependency among frames)
Predicted importance scores
0.52
0.37
0.91
0.97
0.15
Frame importance scores according to users
0.6
0.1
0.8
1
0.2
-

Video Summarization Using Fully Convolutional Sequence Networks



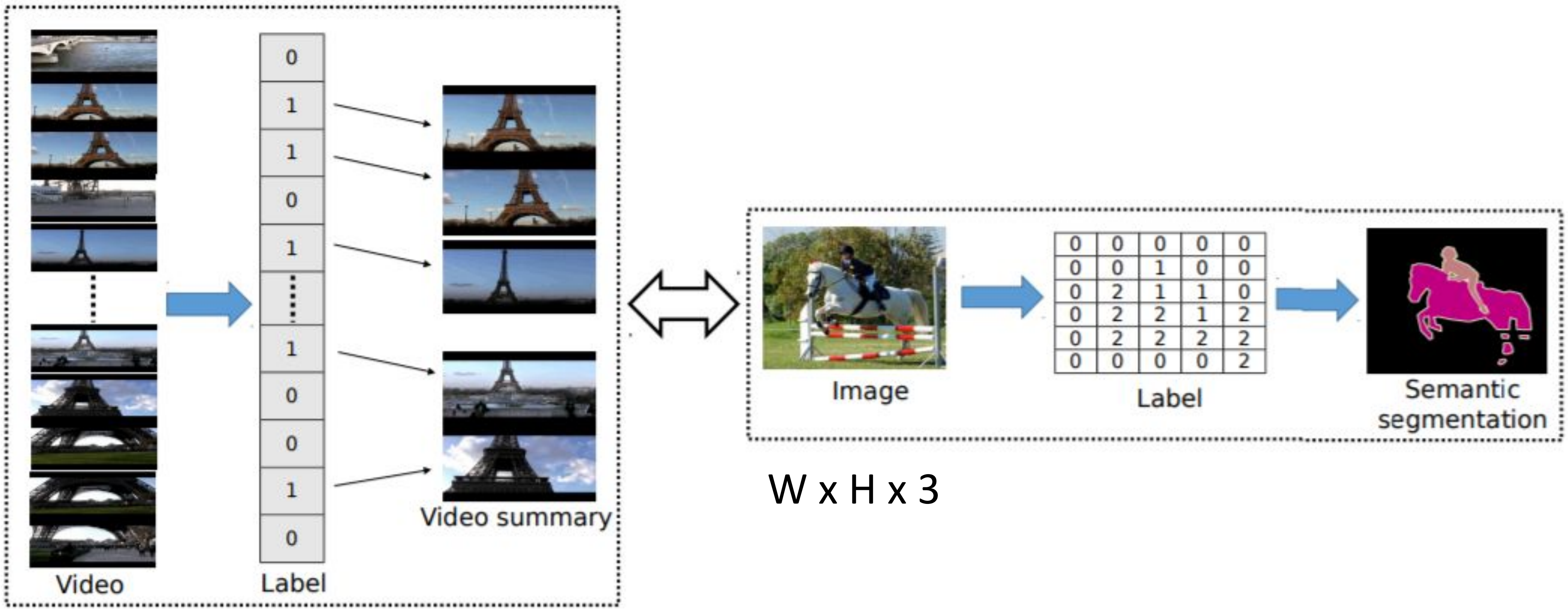
M. Rochan, L. Ye, and Y. Wang, Video Summarization Using Fully Convolutional Sequence Networks (ECCV 2018)

Video Summarization Using Fully Convolutional Sequence Networks



M. Rochan, L. Ye, and Y. Wang, Video Summarization Using Fully Convolutional Sequence Networks (ECCV 2018)

Video Summarization Using Fully Convolutional Sequence Networks



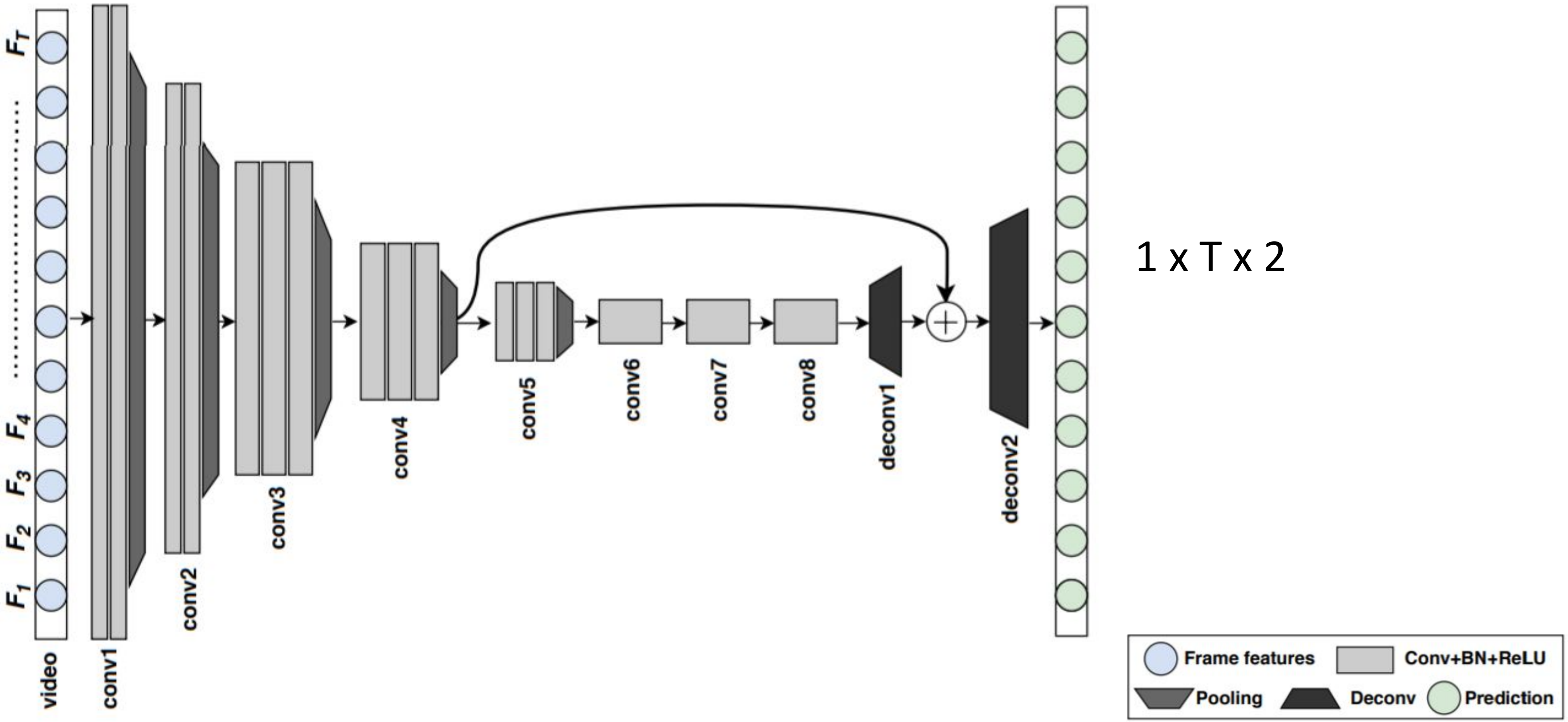
0	0	0	0	0
0	0	1	0	0
0	2	1	1	0
0	2	2	1	2
0	2	2	2	2
0	0	0	0	2

W x H x 3

1 x T x K

M. Rochan, L. Ye, and Y. Wang, Video Summarization Using Fully Convolutional Sequence Networks (ECCV 2018)

Video Summarization Using Fully Convolutional Sequence Networks



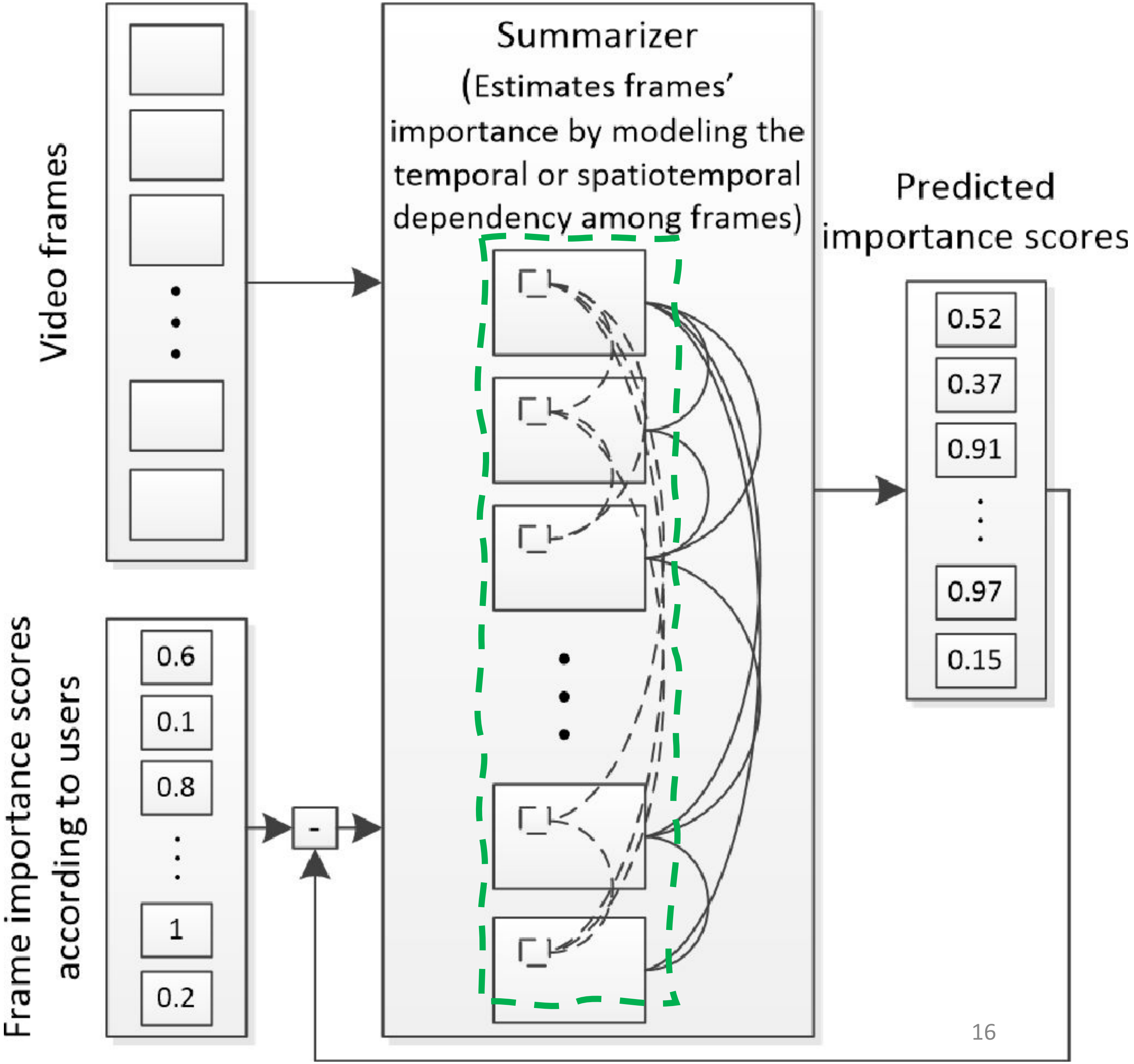
1 x T x 2

M. Rochan, L. Ye, and Y. Wang, Video Summarization Using Fully Convolutional Sequence Networks (ECCV 2018)

Unimodal approaches

Supervised Learning

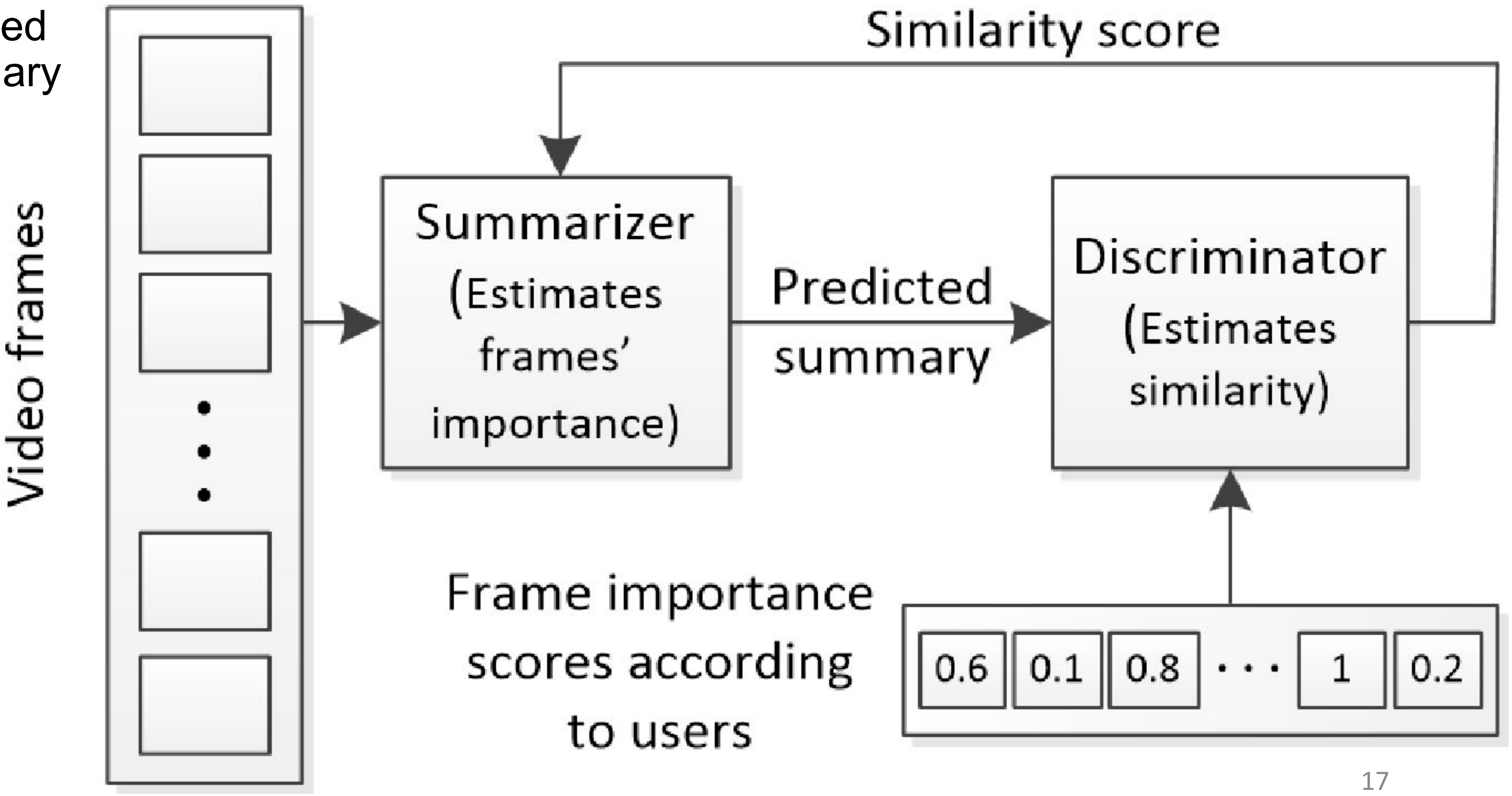
- Learn frame importance by modeling the **spatiotemporal** structure of the video


Video frames
Summarizer
(Estimates frames' importance by modeling the temporal or spatiotemporal dependency among frames)
Predicted importance scores
0.52
0.37
0.91
0.97
0.15
Frame importance scores according to users
0.6
0.1
0.8
1
0.2
-

Unimodal approaches

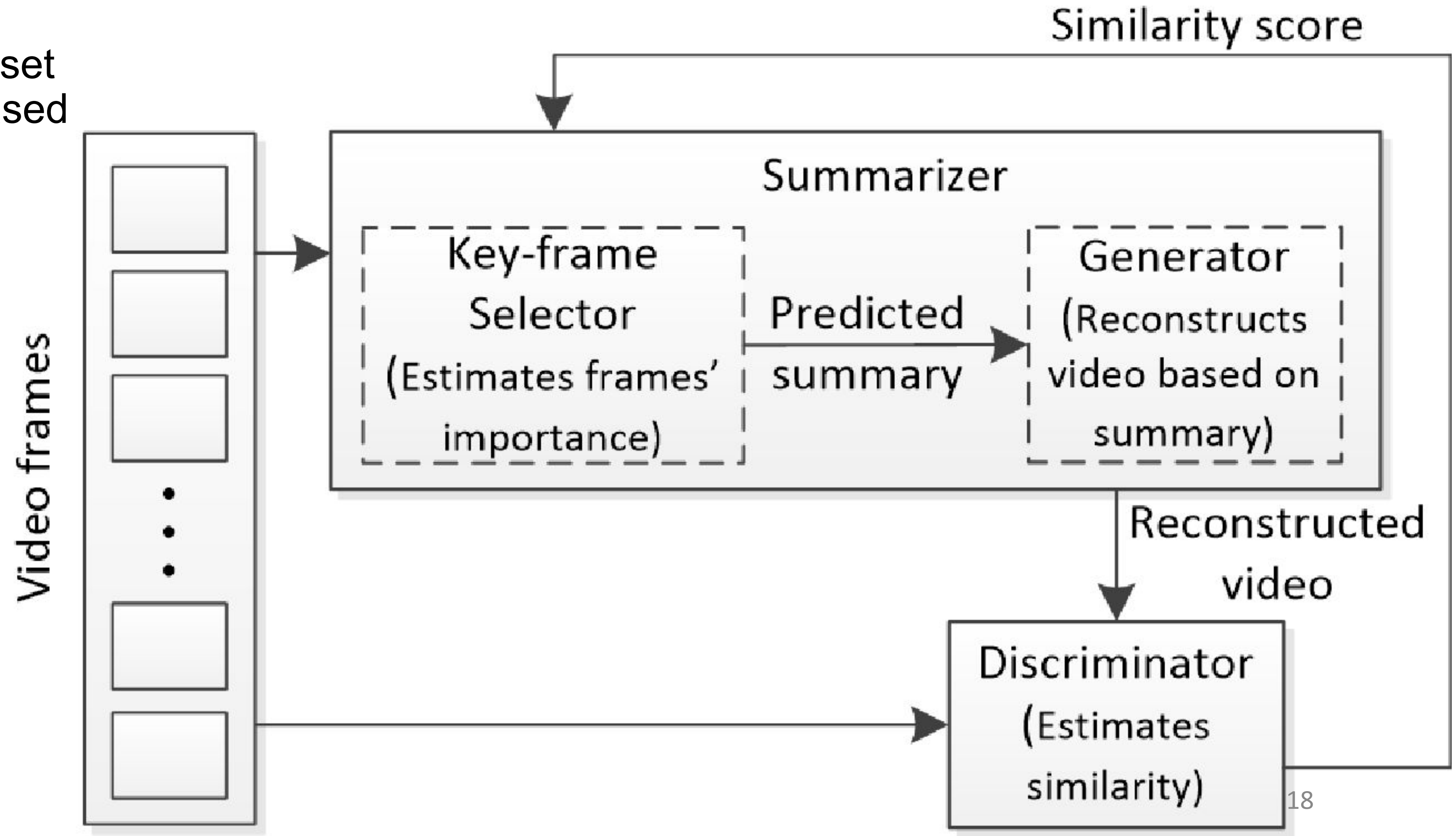
Supervised Learning

- Learn summarization by fooling a **discriminator** when trying to discriminate a machine-generated from a human-generated summary


Video frames
Similarity score
Summarizer (Estimates frames' importance)
Predicted summary
Discriminator (Estimates similarity)
Frame importance scores according to users
0.6
0.1
0.8
1
0.2

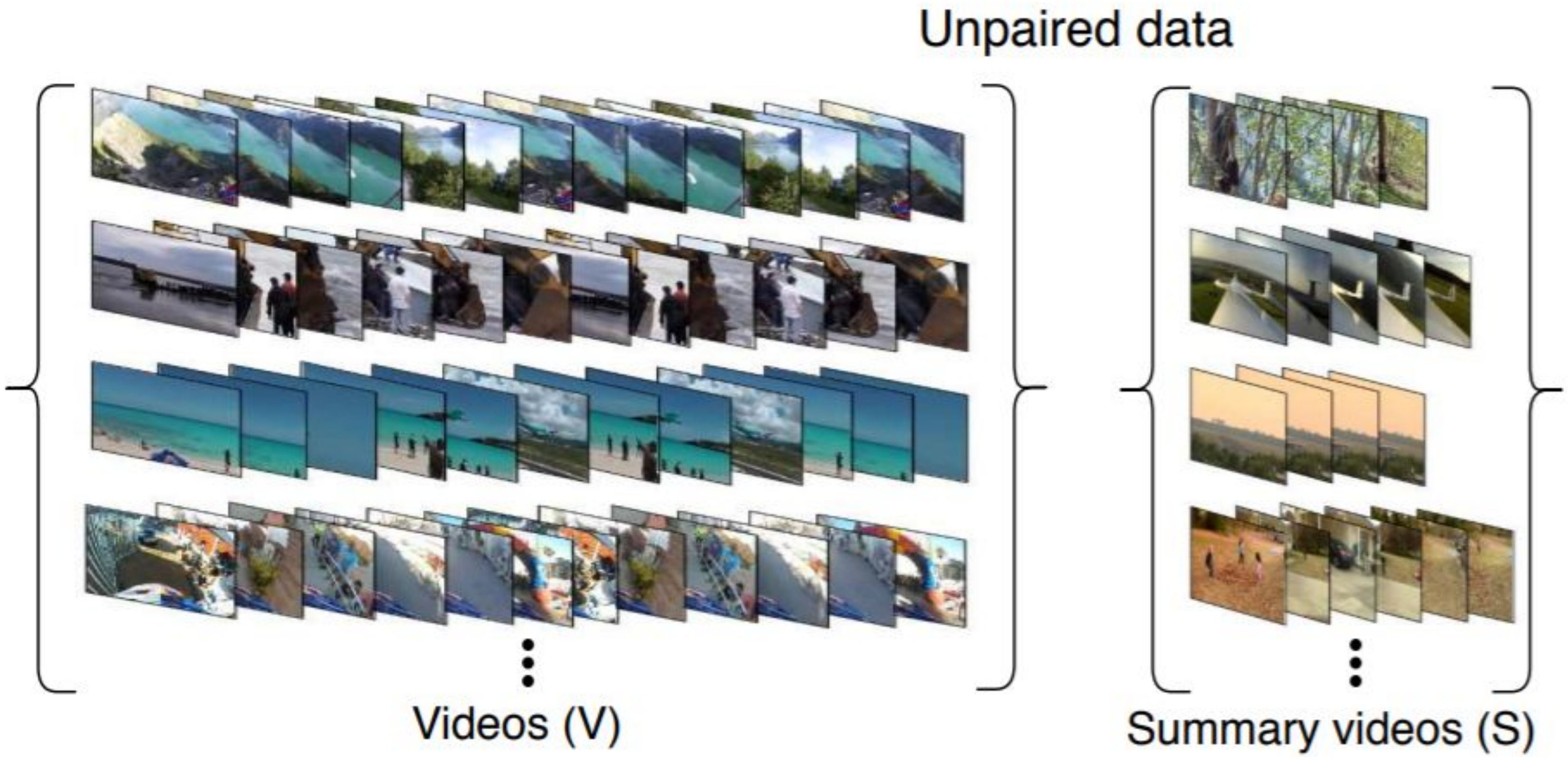
Unimodal approaches
Unsupervised Learning

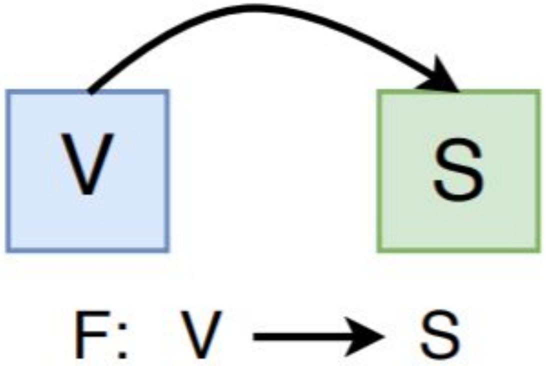
- Learn summarization by fooling a **discriminator** when trying to discriminate the original video (or set of keyframes) from a summary-based reconstruction of it



Similarity score
Video frames
Summarizer
Key-frame Selector (Estimates frames' importance)
Predicted summary
Generator (Reconstructs video based on summary)
Reconstructed video
Discriminator (Estimates similarity)

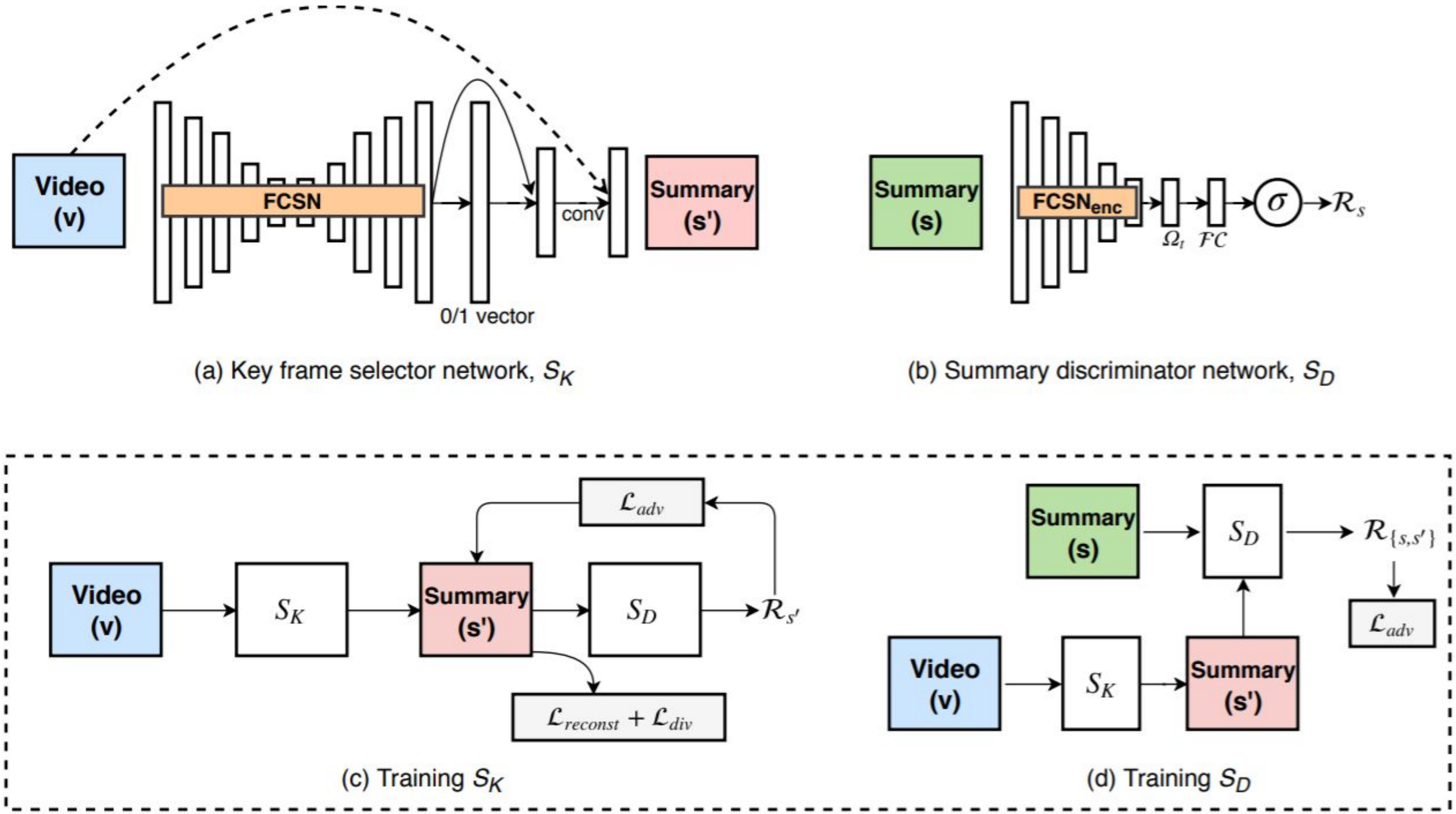
Video Summarization by Learning from Unpaired Data





M. Rochan, and Y. Wang, Video Summarization by Learning from Unpaired Data (CVPR 2019)

Video Summarization by Learning from Unpaired Data



(a) Key frame selector network, S_K

(b) Summary discriminator network, S_D

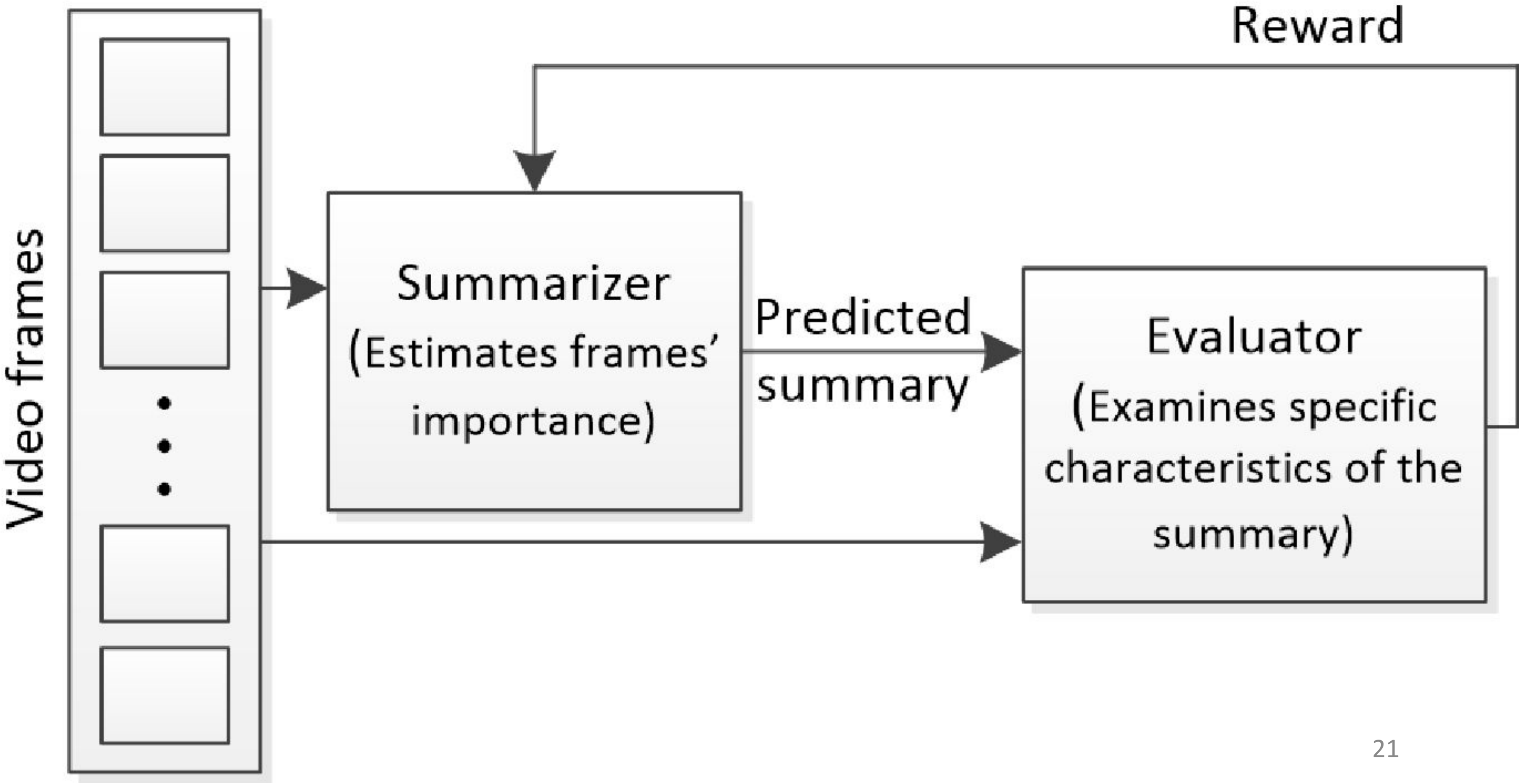
(c) Training S_K

(d) Training S_D

M. Rochan, and Y. Wang, Video Summarization by Learning from Unpaired Data (CVPR 2019)

Unimodal approaches
Unsupervised Learning

- Learn summarization by targeting **specific desired properties** for the summary


Reward
Video frames
Summarizer
(Estimates frames' importance)
Predicted summary
Evaluator
(Examines specific characteristics of the summary)

Unimodal approaches

Unsupervised Learning

- Build object-oriented summaries by modeling the key-motion of important visual objects
    - perform a preprocessing step to find important objects and their key-motions
    - represent the whole video by creating super-segmented object motion clips
    - generate summaries that show the representative objects in the video and the key-motions of each of these objects

Unimodal approaches

Weakly-supervised Learning

- Learn from semantically similar web videos
  - Use video-level metadata to define a categorization of videos.
  - Leverage multiple videos of a category to extract features and learn to automatically categorize new videos.
  - Use the learned model to select the video segments that maximize the relevance between the summary and the video category.

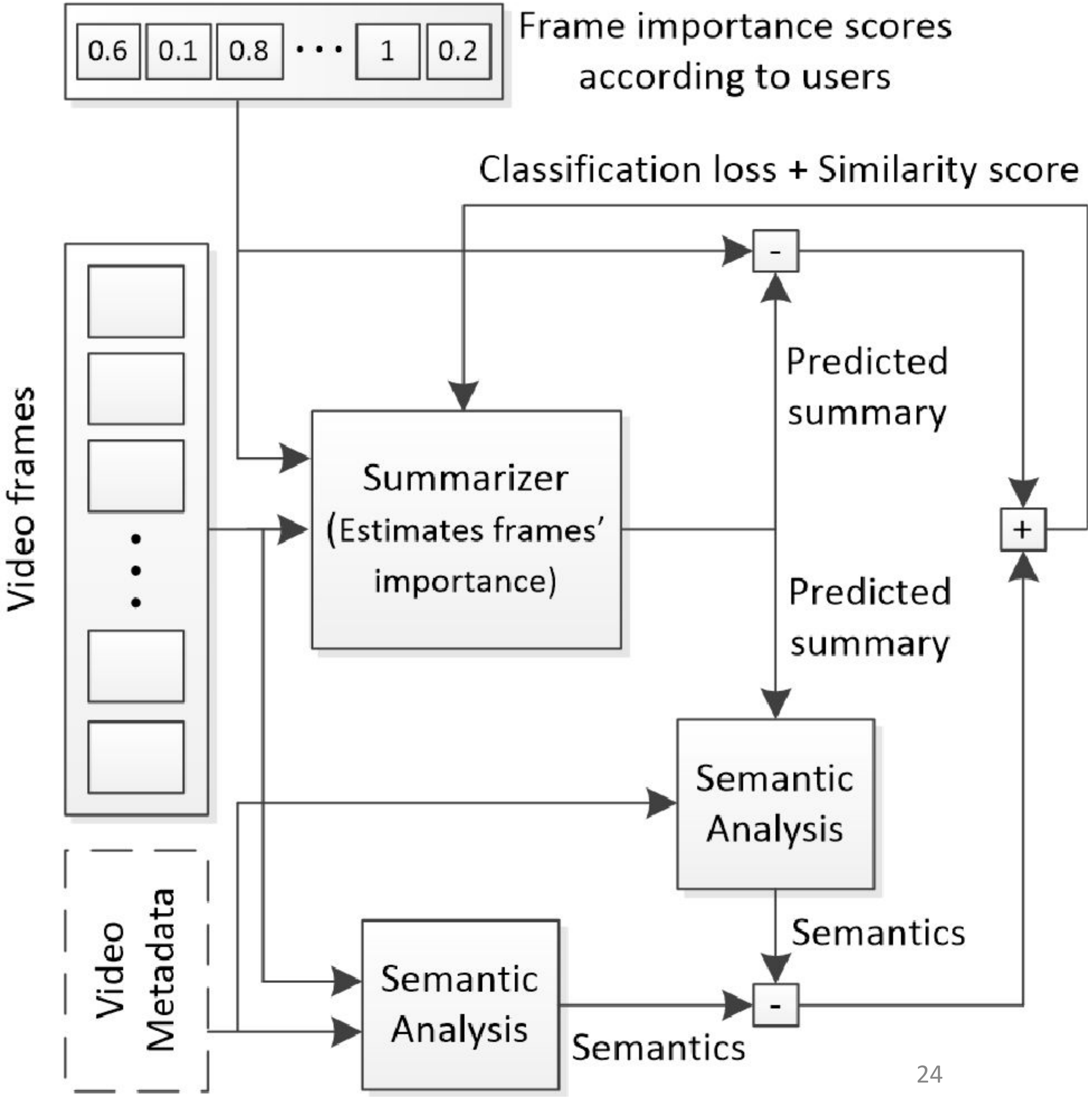
- Learn using annotations from a similar domain
  - Learn from third-person annotated videos.
  - Exploit transfer learning to learn how to summarize first-person videos.

- Learn using weakly/sparsely-labeled data
  - Typically use reinforcement learning

Multimodal approaches

Supervised Learning

- Use textual video metadata
- Use other types of data



Current state of development: Supervised Approaches

- The best-performing supervised approaches utilize tailored attention mechanisms (to capture variable-range temporal dependency) or memory networks (to capture long-range temporal dependencies).
- Some works exhibit high performance in one of the datasets and very low or random performance in the other datasets (indication of overfitting).
- Multimodality approaches are not yet competitive compared to the unimodal ones that rely on the analysis of the visual content only.
- The use of weak labels does not yet enable good summarization.

Current state of development: Unsupervised Approaches

- The use of GANs seems to be the most promising choice, as GAN-based methods perform the best among unsupervised approaches.

- The use of attention mechanisms helps to identify the important parts of the video and boost performance.

- Techniques that rely on reward functions and reinforcement learning are not yet competitive compared to GAN-based methods.

- Some methods low or random performance.

Future Directions

- Major research direction is towards the development of supervised algorithms.
- Unsupervised video summarization methods that combine the merits of adversarial and reinforcement learning should be further explored.
- Advanced multi-head attention mechanisms, for better estimating variable-range temporal dependencies among parts of the video.
- Extend LSTM architectures with high-capacity memory networks, to capture long-range dependencies of the visual content, especially for long videos (e.g., movies).
- Introduce domain-specific rules in the unsupervised video summarization process (i.e., introducing the human in the loop).
- Multimodal summarization approaches using both visual and audio modality of the video, consider audio segmentation to produce more natural story narration.