Master in
Computer Vision
Barcelona

Module 6: Video Analysis
Lecture 9: Self-supervised and multimodal learning
Lecturer: Gloria Haro

What is Self-Supervised Learning?

Self-Supervised Learning (SSL) is a form of unsupervised learning where **the data provides the supervision**.

The network learns from unlabeled data.

How? A proxy loss is defined and the network learns a useful representation in order to solve it.

Motivation for SSL

Image source: https://www.oberlo.com/blog/youtube-statistics

Motivation for SSL

- Data annotation is expensive and time-consuming, especially in video!

Image source: https://www.oberlo.com/blog/youtube-statistics

Motivation for SSL

- Data annotation is expensive and time-consuming, especially in video!
- Avoid ambiguities or biases in human-generated annotation

Image source: https://www.oberlo.com/blog/youtube-statistics

Motivation for SSL

- Data annotation is expensive and time-consuming, especially in video!
- Avoid ambiguities or biases in human-generated annotation
- Tones of unlabelled data

Image source: https://www.oberlo.com/blog/youtube-statistics

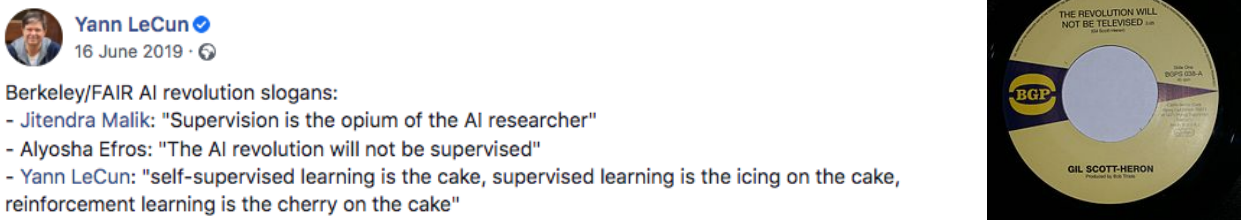
Motivation for SSL

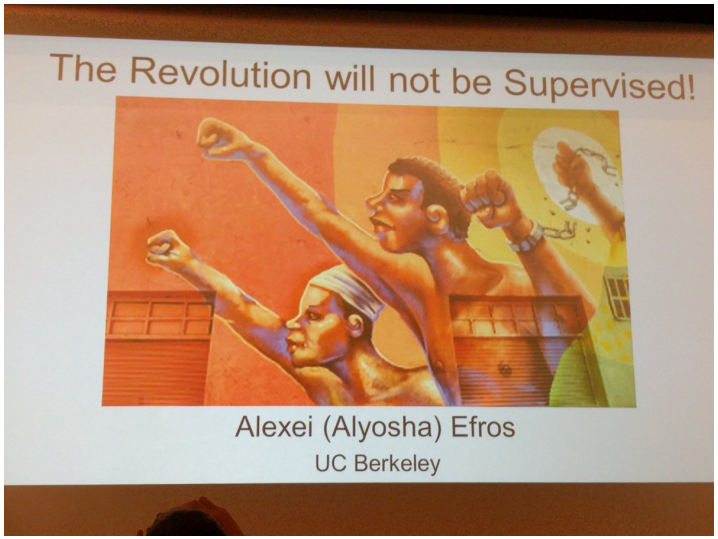
- Data annotation is expensive and time-consuming, especially in video!
- Avoid ambiguities or biases in human-generated annotation
- Tones of unlabelled data

- How infants partly learn

Image source: https://www.oberlo.com/blog/youtube-statistics

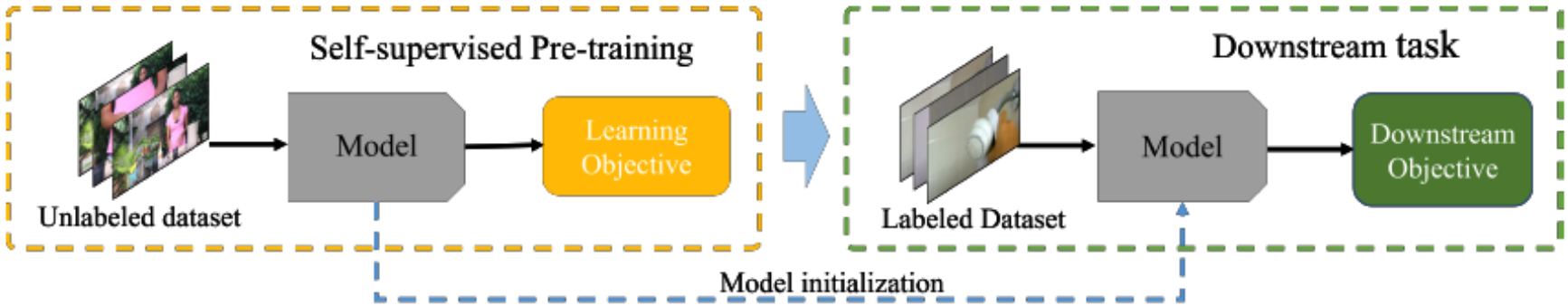
Motivation for SSL





With thanks to Alyosha Efros and Gil Scott Heron

Typical use of SSL



M.C. Schiappa, Y.S. Rawat, M. Shah. Self-supervised learning for videos: A survey. ACM Computing Surveys 2022.

Image source: [Schiappa et al. 2022]

SSL in video vs SSL in images

Learning video representations is **more challenging** due to the extra temporal dimension.

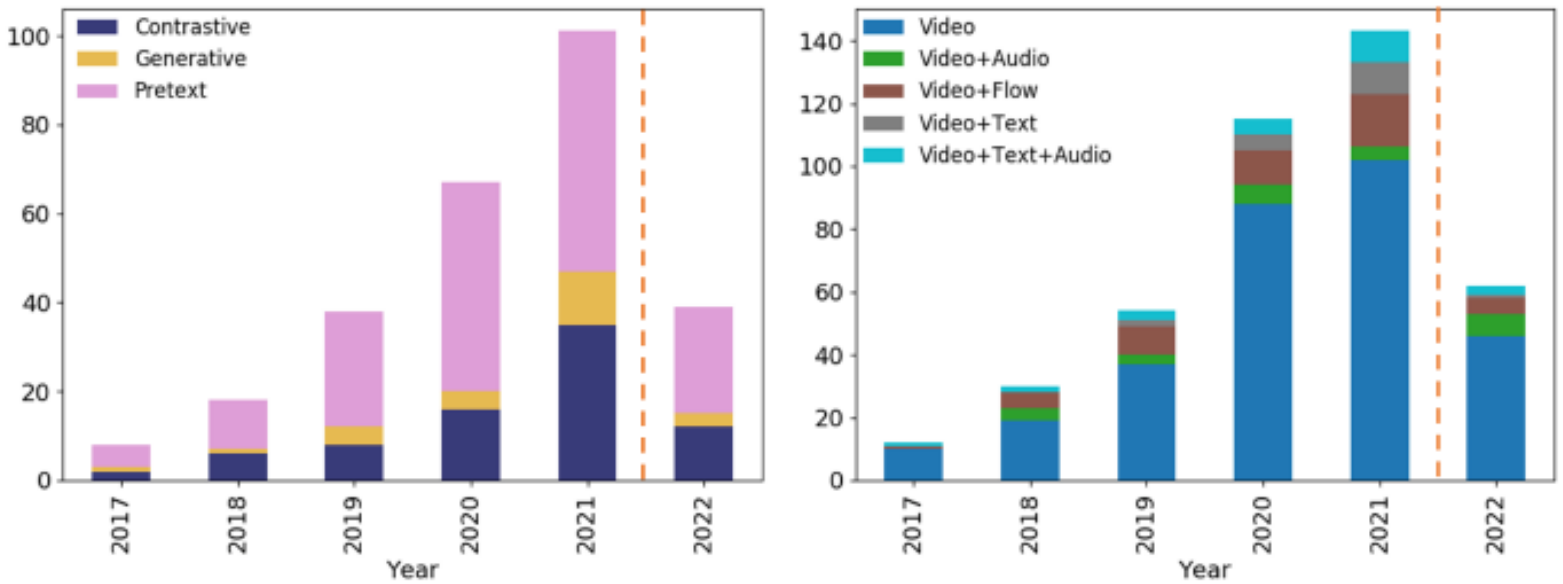
However, video provides **new opportunities** for video-exclusive ideas, thanks to, e.g.

- Temporal coherence
- Appearance variations of objects over time
- Strong correlation between audio and video modalities

Some approaches are extensions/inspirations to/from SSL techniques for images.

Classification of SSL techniques in video

- Pretext learning
- Generative learning
- Contrastive learning
- Cross-modal agreement



NOTE: The year 2022 remains incomplete because a majority of the conferences occur later in the year.

M.C. Schiappa, Y.S. Rawat, M. Shah. Self-supervised learning for videos: A survey. ACM Computing Surveys 2022.

Image source: [Schiappa et al. 2022]

Outline

- Introduction to SSL
- SSL approaches in video
  - **Pretext learning**
  - Generative learning
  - Contrastive learning
  - Cross-modal agreement
- Multimodal and cross-modal approaches

Pretext learning

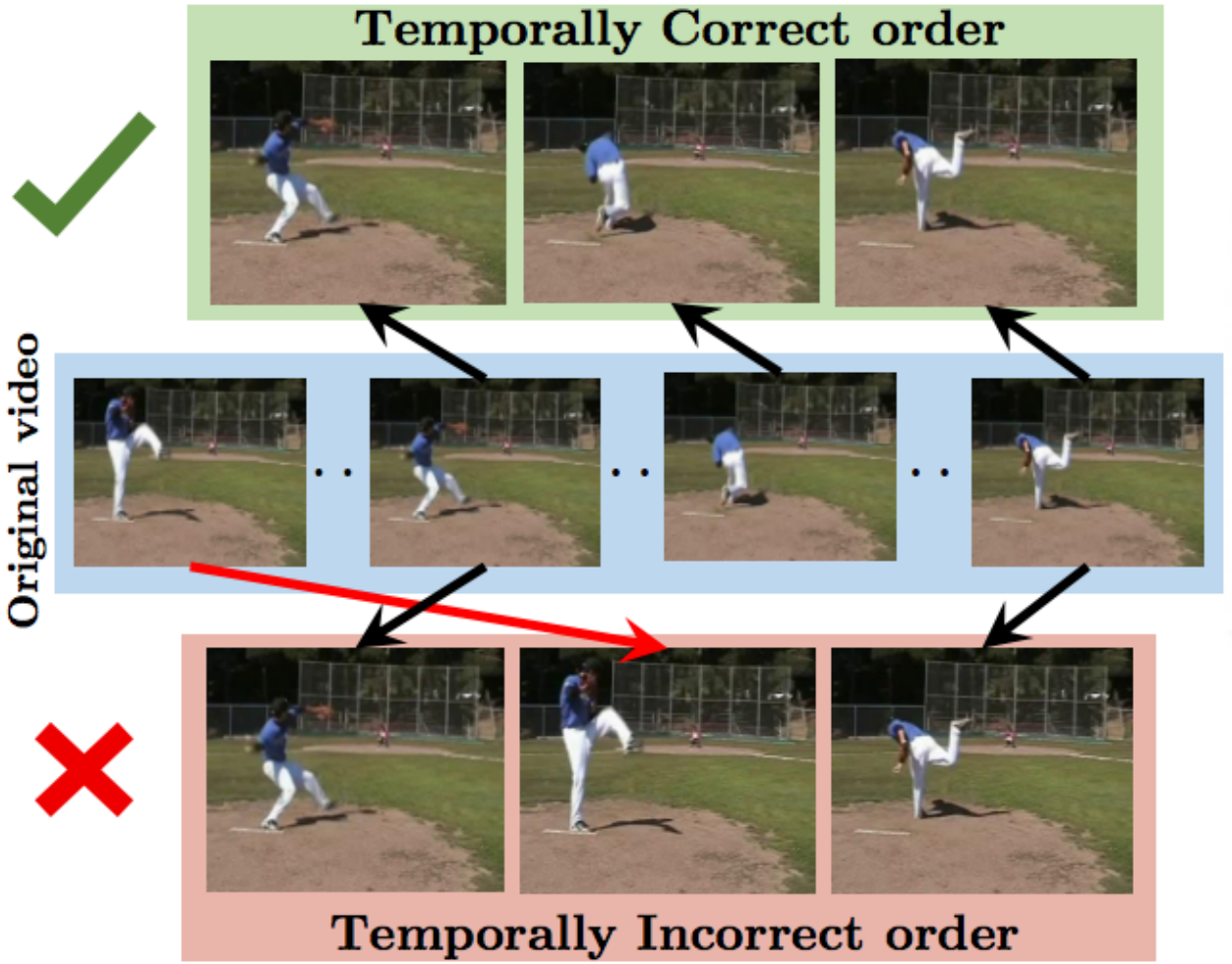
A **pretext or surrogate task** is defined and the network is trained to solve it.

The surrogate task doesn't need manual annotation.

The task at hand should require a high-level understanding of the input data in order to solve it. $\rightarrow$ **As a by-product, the network learns generalizable features.**

Pretext learning

Sequential verification task



I. Misra, C. L. Zitnick, M. Hebert. Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification. ECCV 2016.

Image source: [Misra et al. 2016]

Pretext learning

Sequential verification task

Determine wether a sequence of video frames are in the correct temporal order → **Binary classification task**

Solving the task requires reasoning about object trasnformations and relative locations through time. → Forces the representation to capture object appearances and deformations.

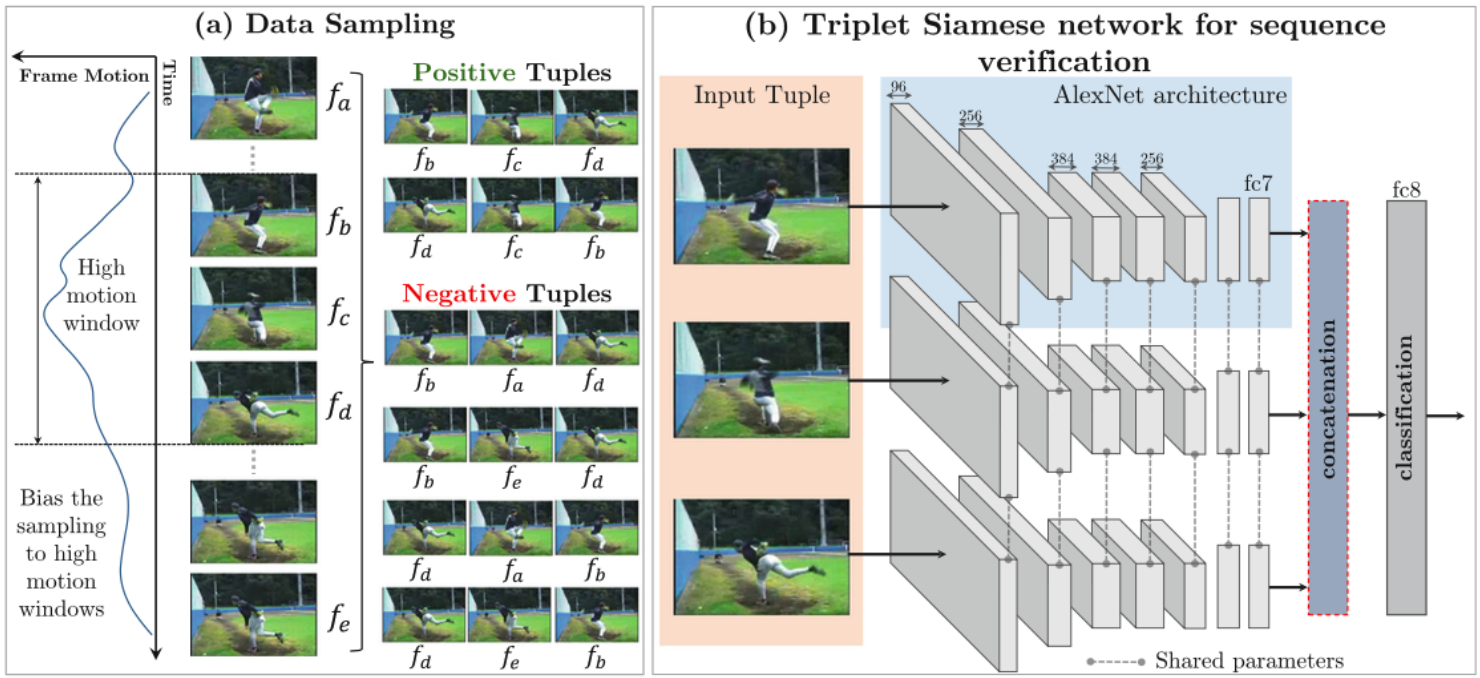


Image source: [Misra et al. 2016]

Pretext learning

Sequential verification task

Focus on videos with human actions

Learned visual features useful for the **downstream tasks**:

- Action recognition
- Pose estimation

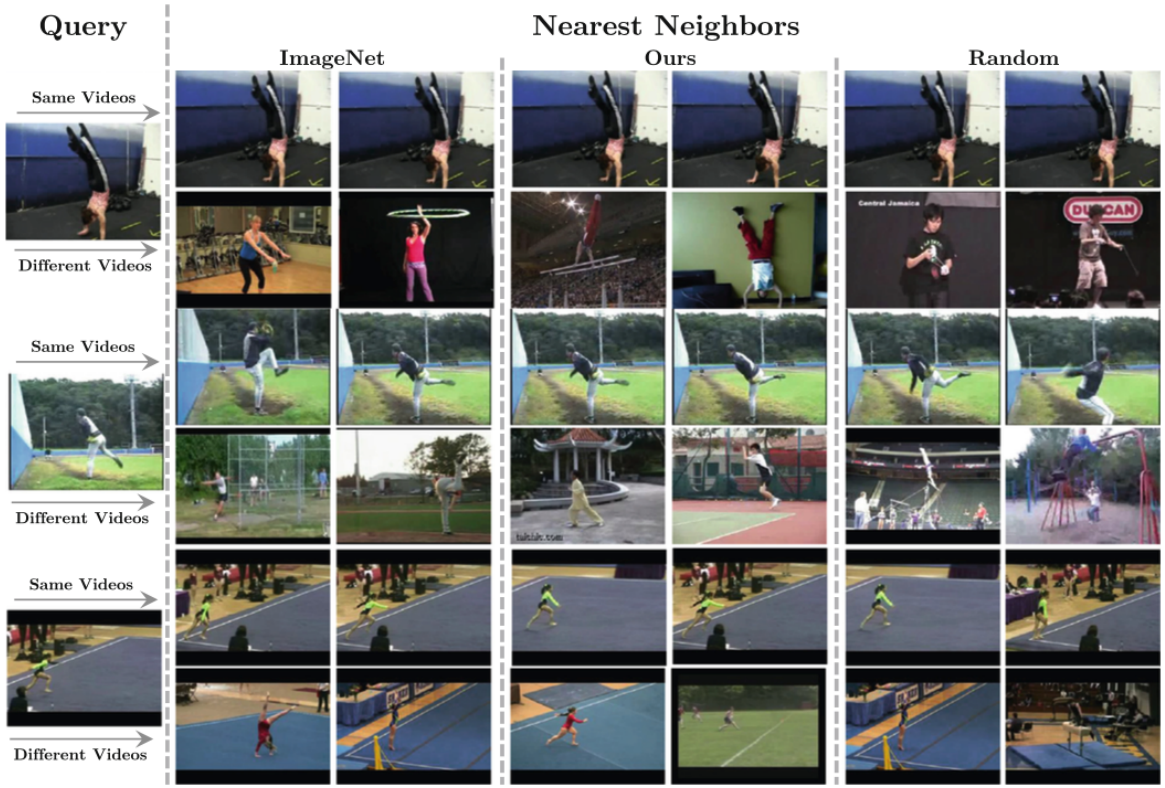
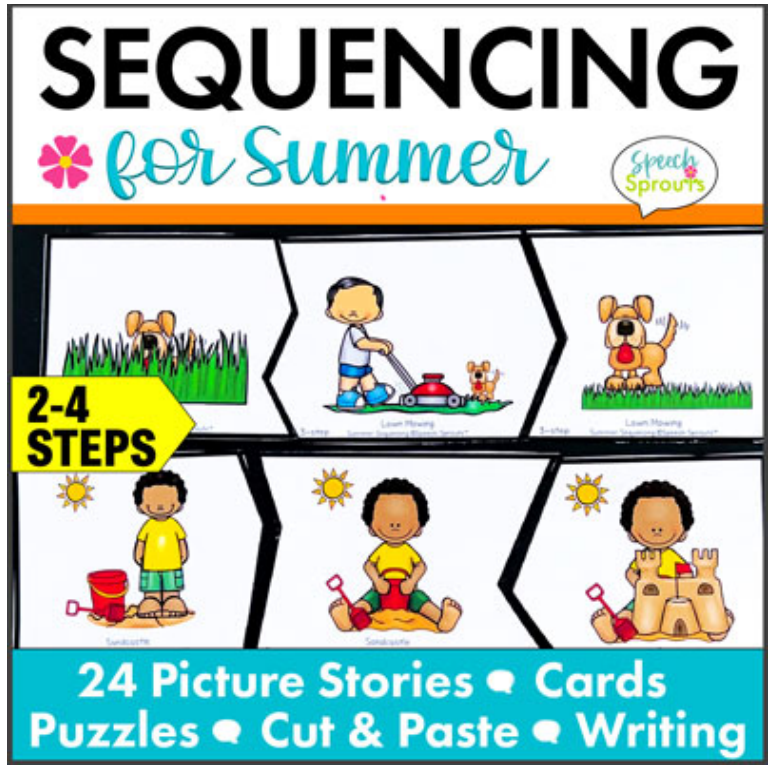


Image source: [Misra et al. 2016]

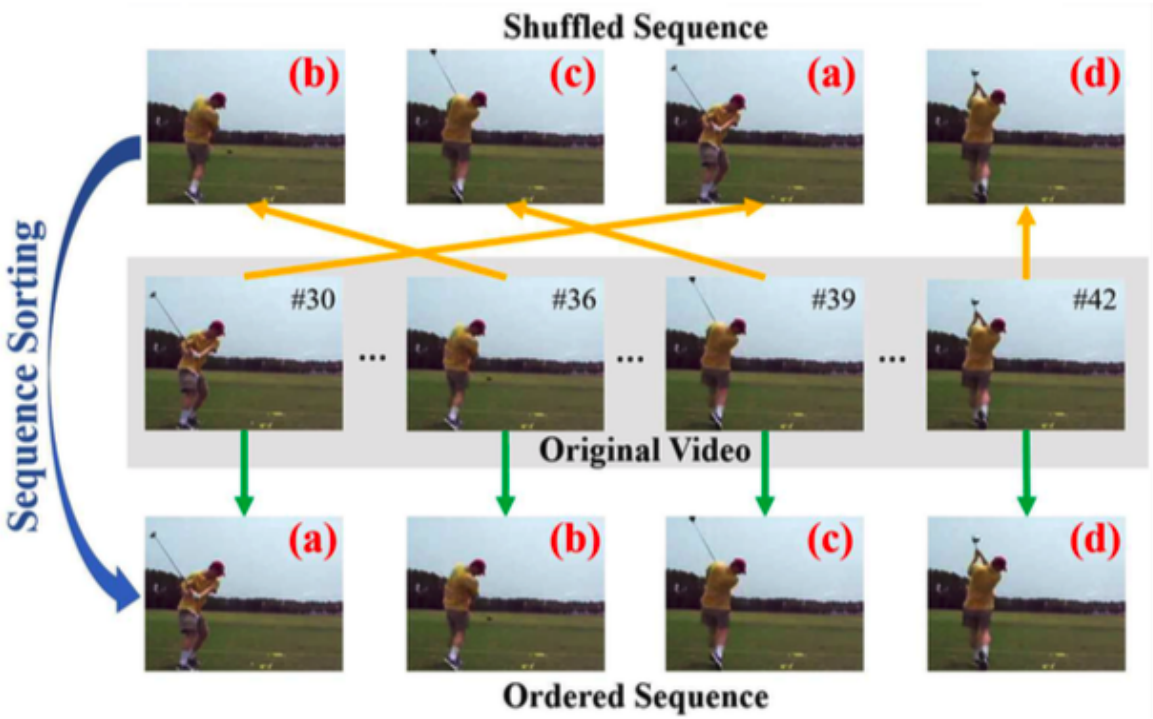
Pretext learning



Images source: Amazon and https://www.teacherspayteachers.com/

Pretext learning

Sequence sorting task



The supervisory signal is richer than in the previous case.

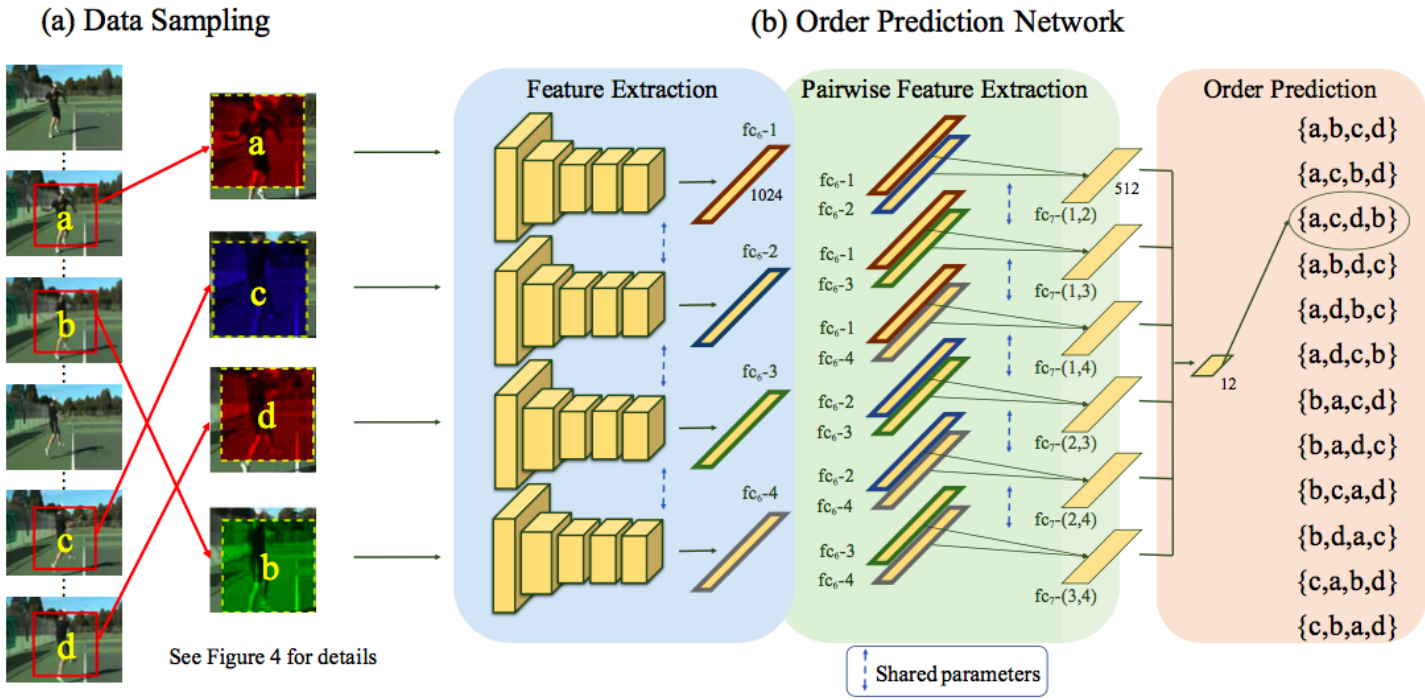
The network has to predict the correct order of frames
$\rightarrow$ **Multi-class classification task**

H.-Y. Lee, J.-B. Huang, M. Singh, M.-H. Yang. Unsupervised representation learning by sorting sequences. ICCV 2017.

Image source: [Lee et al. 2017]

Pretext learning

Sequence sorting task



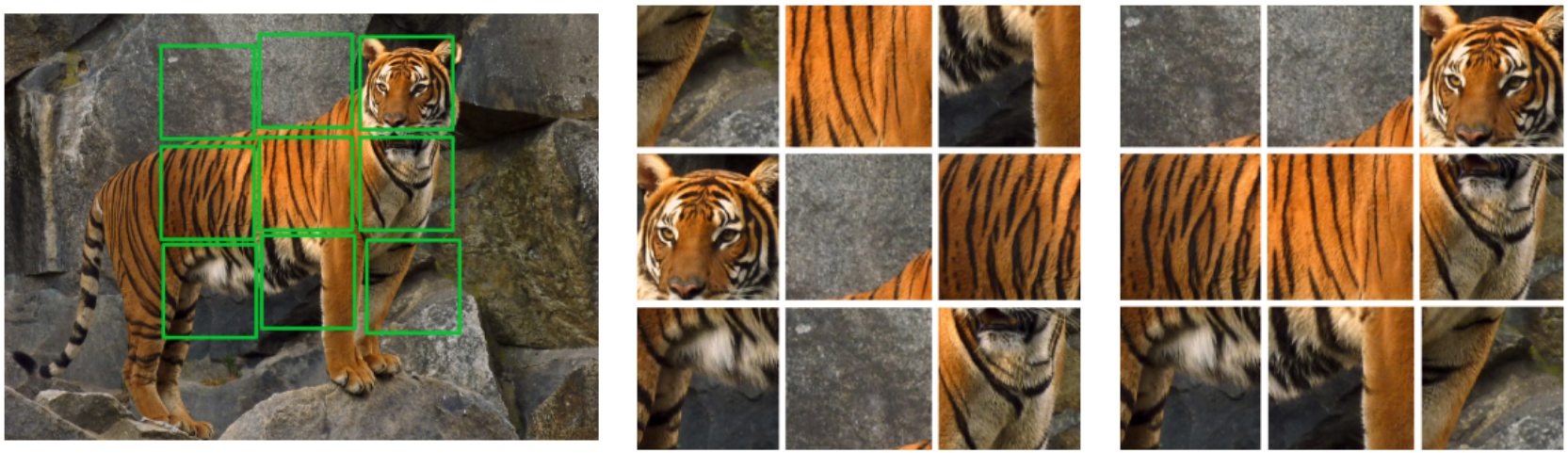
Key aspects:

- Sampling candidate 4-tuples based on motion magnitude
- Random spatial jittering and channel splitting to guide the network to focus on semantics and not on low-level features.
- Pairwise feature extraction

Image source: [Lee et al. 2017]

Pretext learning

Solve a jigsaw puzzle (image)



Downstream tasks: Classification, detection and semantic segmentation.

M.i Noroozi, P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. ECCV 2016.

Image source: [Noroozi and Favaro et 2016]

Pretext learning

Solve a jigsaw puzzle (video)

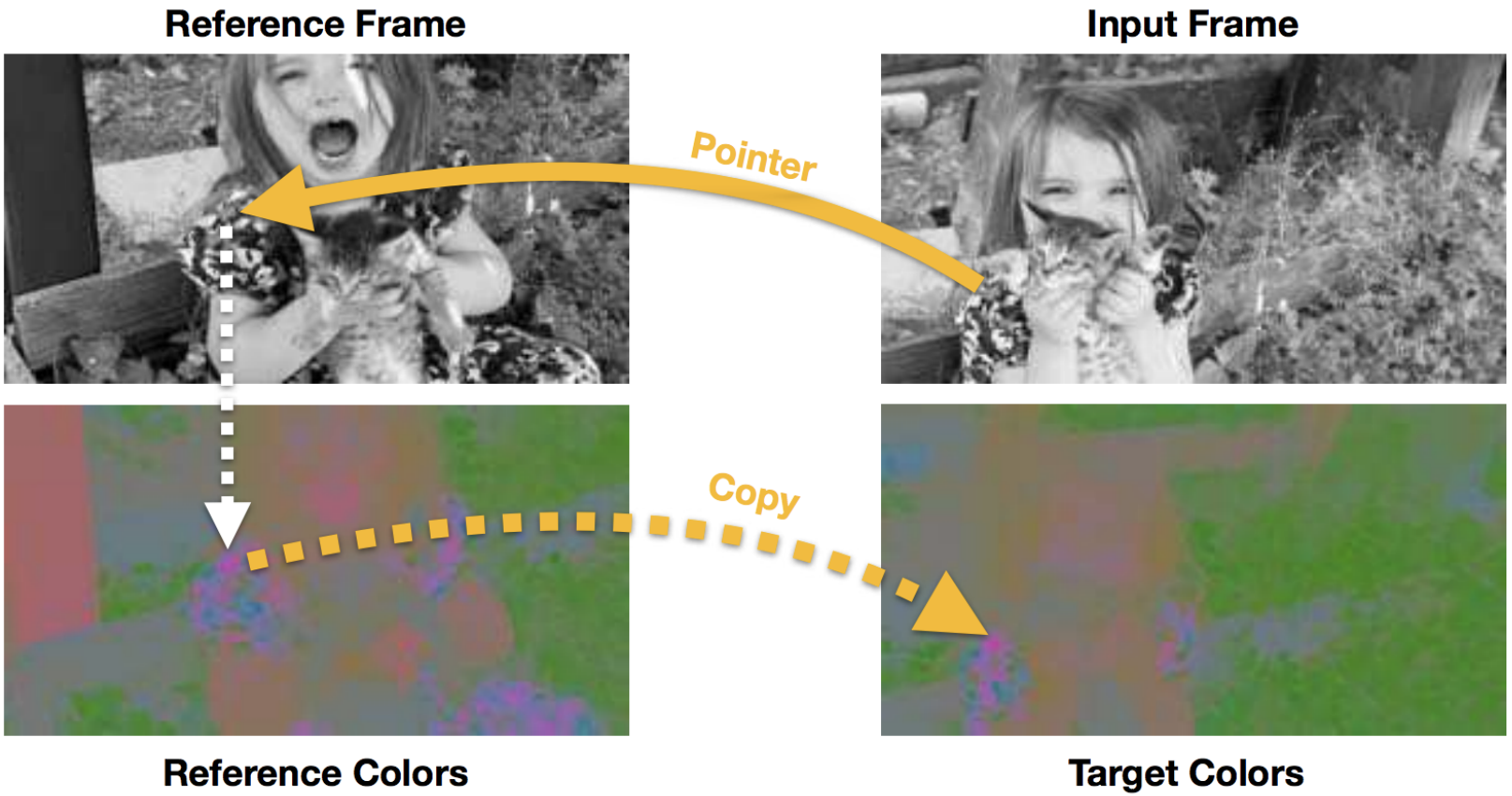
Downstream tasks: Activity recognition

U. Ahsan, R. Madhok, I. Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. WACV 2019.

Image source: [Ahsan et al. 2019]

Pretext learning

Colorization task

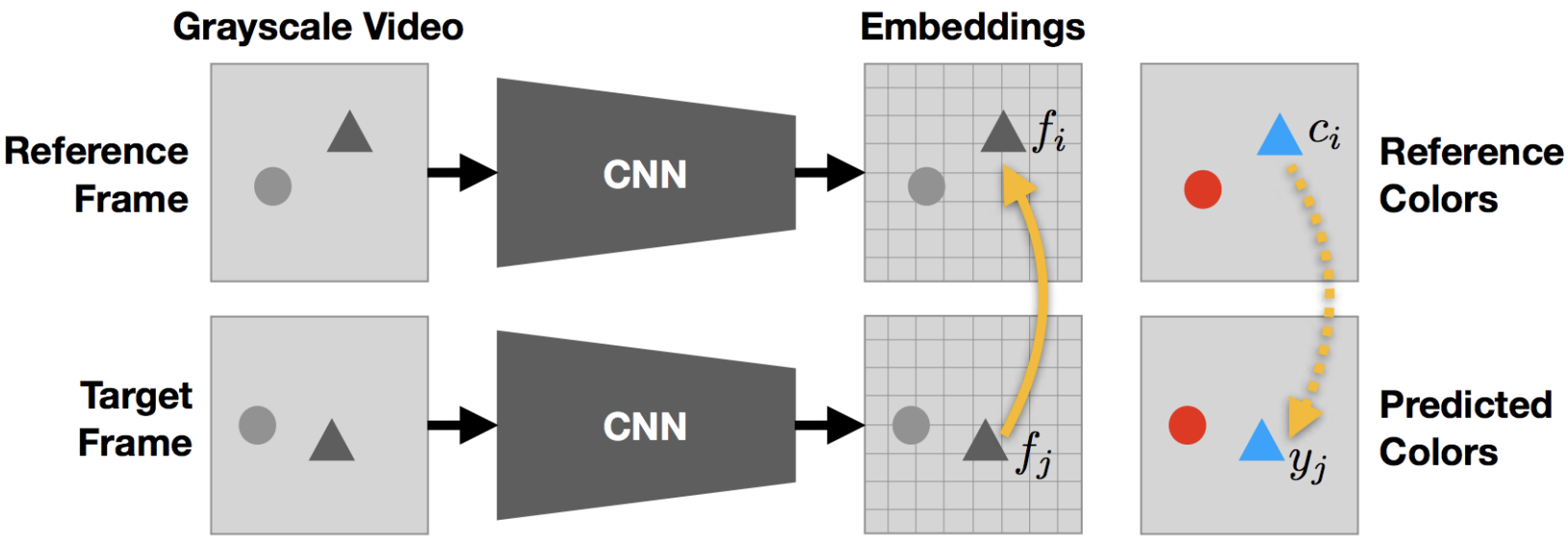


C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, K. Murphy. Tracking emerges by colorizing videos. ECCV 2018.

Image source: [Vondrick et al. 2018]

Pretext learning

Colorization task



Predicted color: $y_j = \sum_i A_{ij} c_i$

A is a **similarity matrix** with elements $A_{ij} = \frac{exp(f_i^T f_j)}{\sum_k exp(f_k^T f_j)}$

Image source: [Vondrick et al. 2018]

Pretext learning

Colorization task

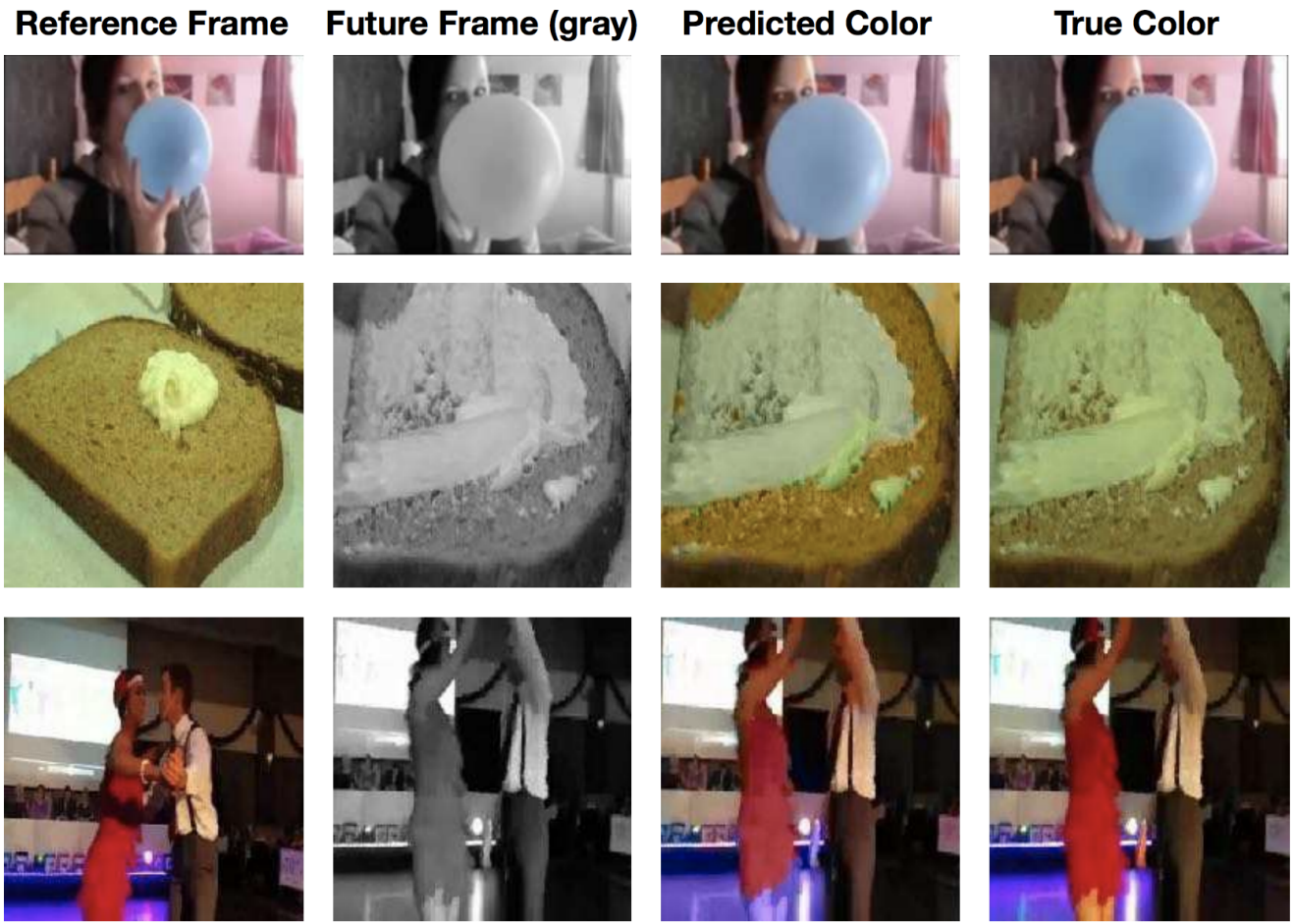


Image source: [Vondrick et al. 2018]

Pretext learning

Self-supervised tracking by colorization

Image source: [Vondrick et al. 2018]

Pretext learning

Self-supervised tracking by colorization

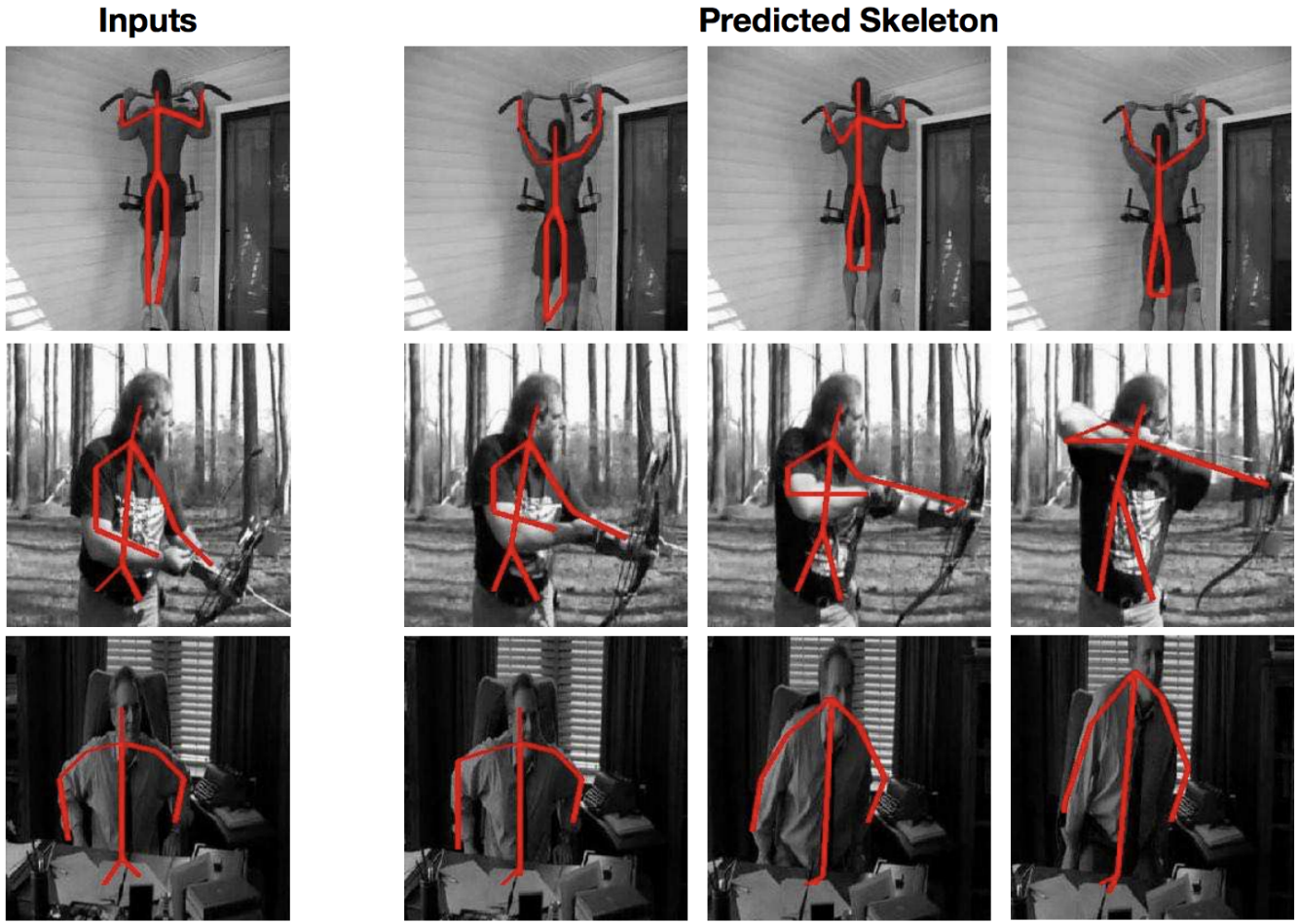
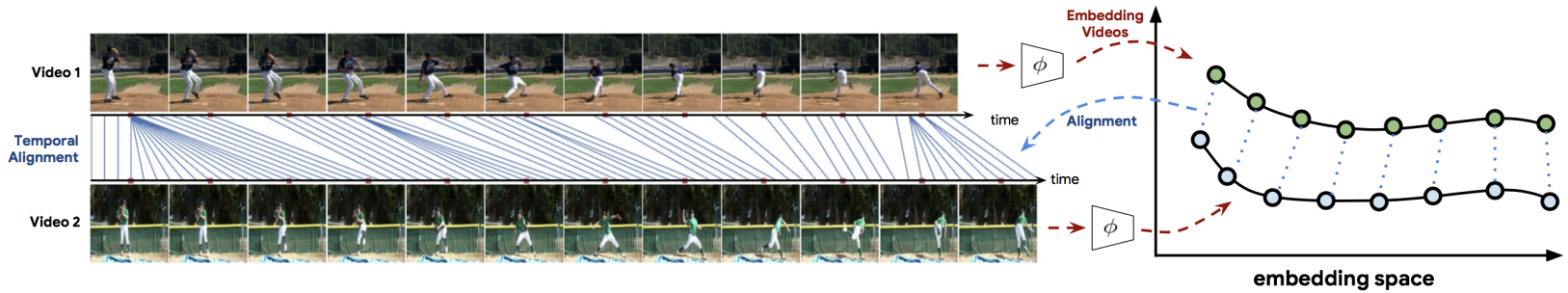


Image source: [Vondrick et al. 2018]

Pretext learning

Temporal alignment between videos



Multiple unaligned videos of the same activity

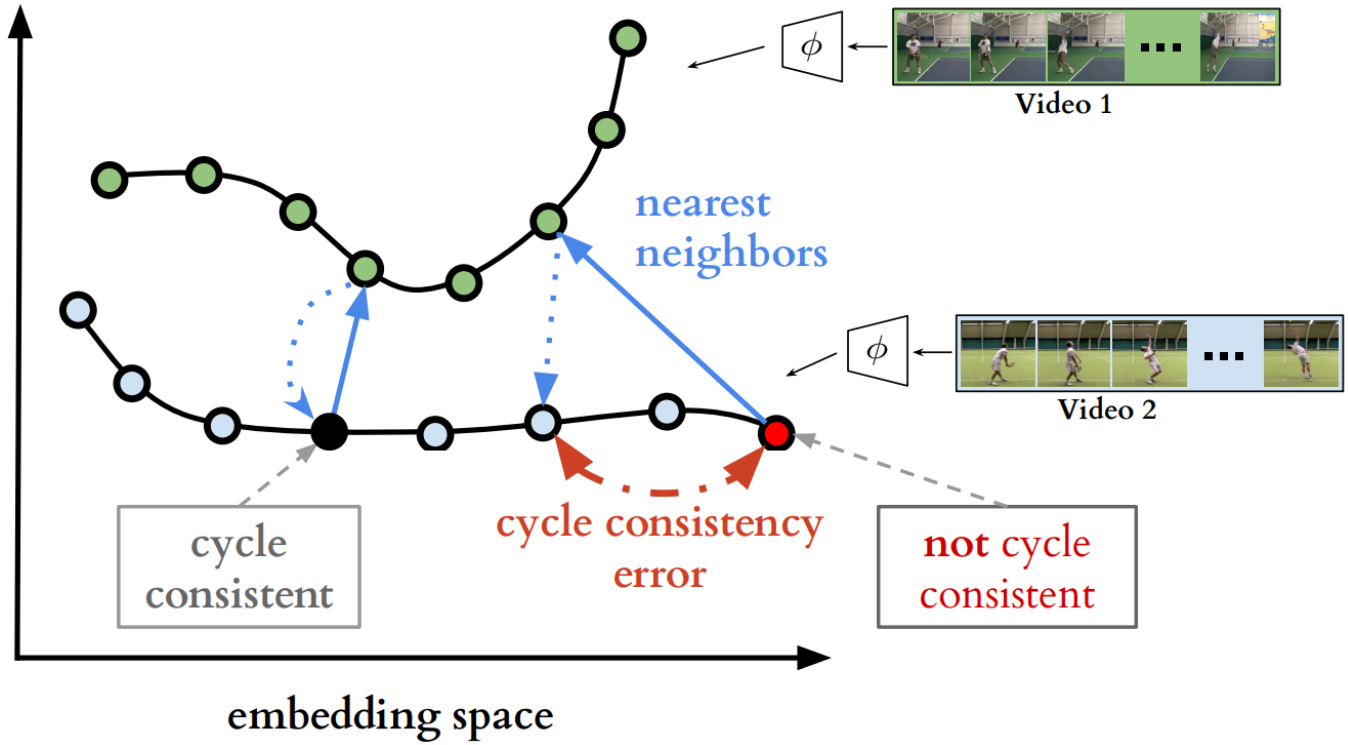
- from different viewpoints
- with different pace (length)
- with different objects
- with camera motion

D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, A. Zisserman. Temporal cycle-consistency learning. CVPR 2019.

Slide source: [Dwibedi et al. 2019]

Pretext learning

Temporal alignment between videos



Goal: Find correspondences across multiple videos despite many factors of variation.

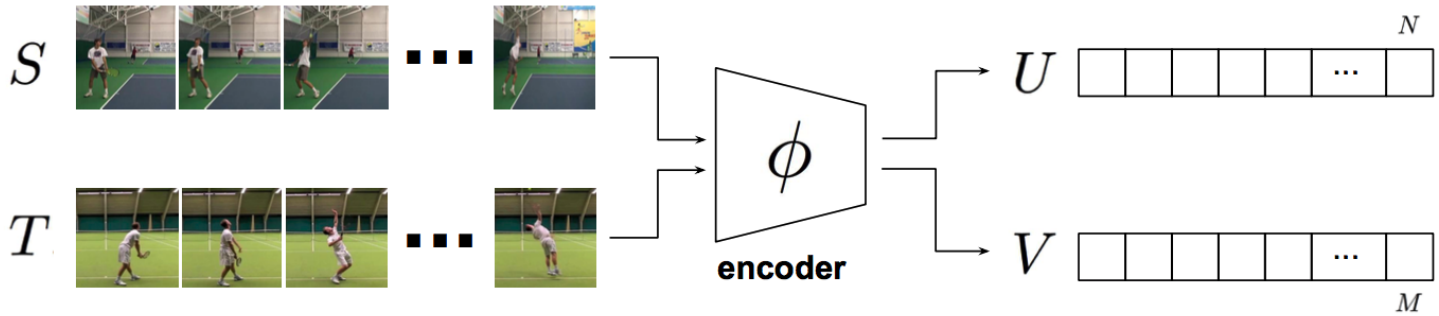
How? Find an embedding space where nearest neighbors give the correspondences

Pretext task: **Temporal cycle consistency**

Image source: [Dwibedi et al. 2019]

Pretext learning

Temporal alignment between videos



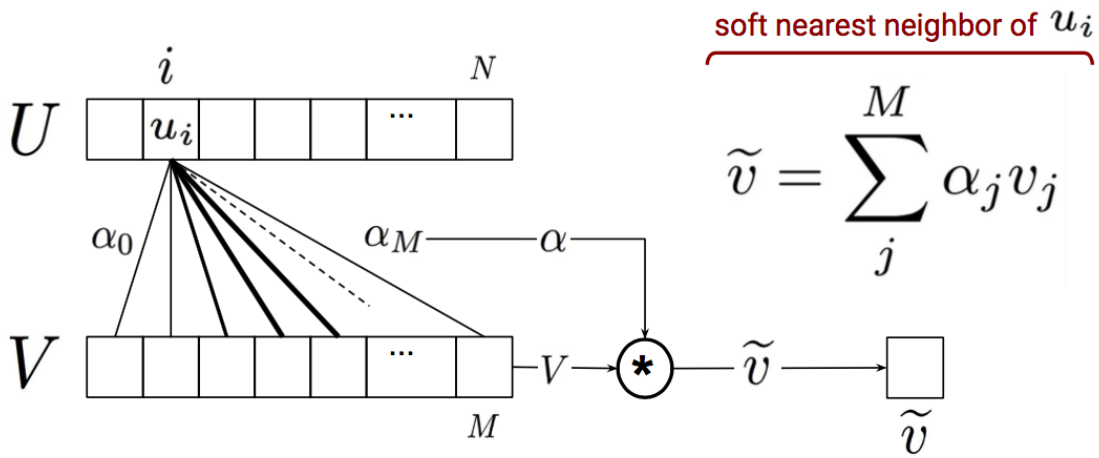
$$U = \{u_1, u_2, ..., u_N\} \qquad V = \{v_1, v_2, ..., v_M\}$$

Slide source: [Dwibedi et al. 2019]

Pretext learning

Temporal alignment between videos

$$\alpha_j = \underbrace{\frac{e^{-||u_i - v_j||_2}}{\sum_k^M e^{-||u_i - v_k||_2}}}_{\text{similarity distribution for } u_i}$$

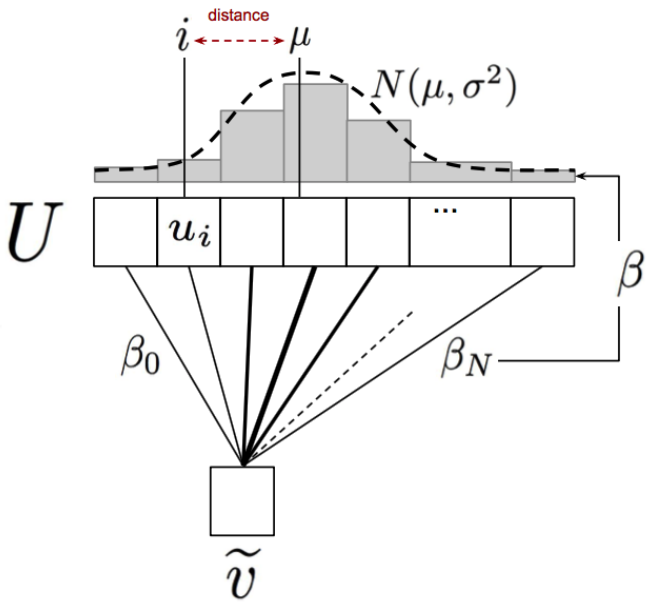


Slide source: [Dwibedi et al. 2019]

Pretext learning

Temporal alignment between videos

$$\beta_k = \underbrace{\frac{e^{-||\widetilde{v}-u_k||_2}}{\sum_j^N e^{-||\widetilde{v}-u_j||_2}}}_{\text{similarity distribution for } \widetilde{v}}$$



Objective Function:

$$L_{cbr} = \frac{|i-\mu|^2}{\sigma^2} + \lambda \log(\sigma)$$

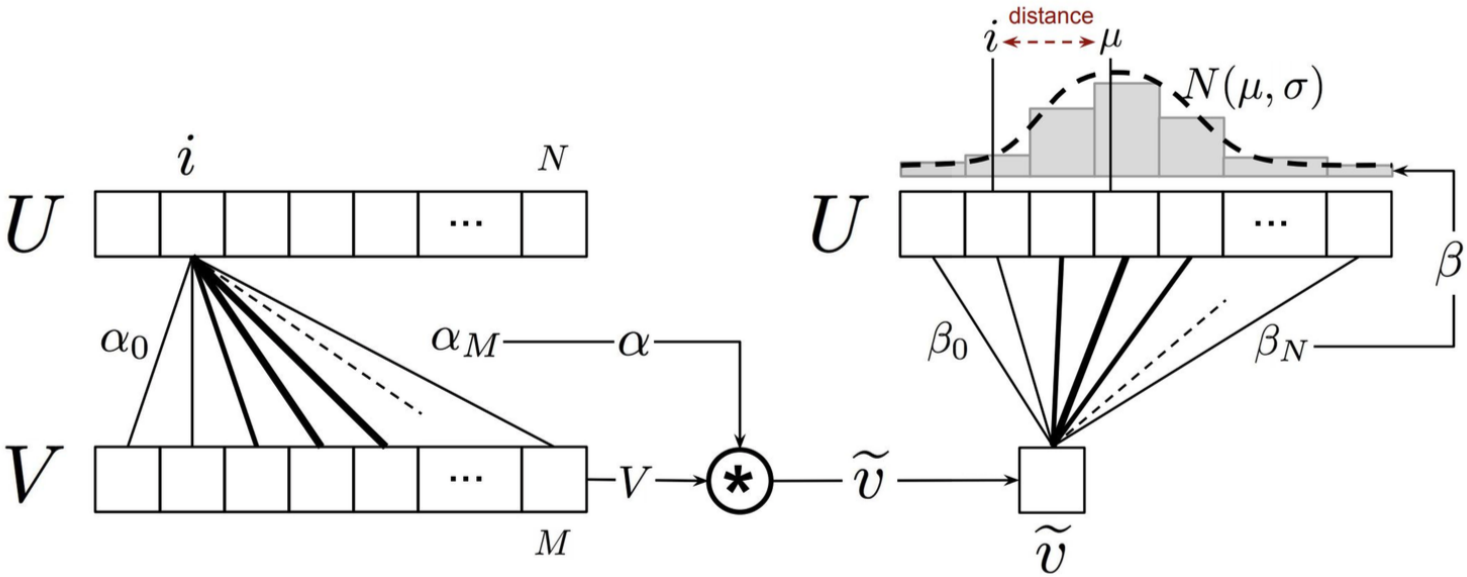
$$\mu = \sum_k^N \beta_k * k$$

$$\sigma^2 = \sum_k^N \beta_k * (k-\mu)^2$$

Slide source: [Dwibedi et al. 2019]

Pretext learning

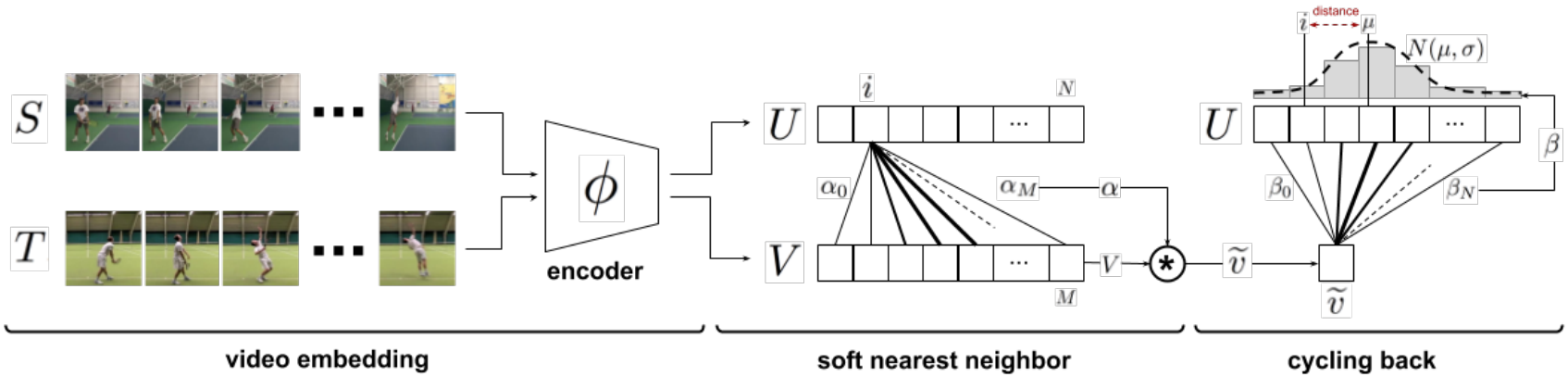
Temporal alignment between videos



Slide source: [Dwibedi et al. 2019]

Pretext learning

Temporal alignment between videos



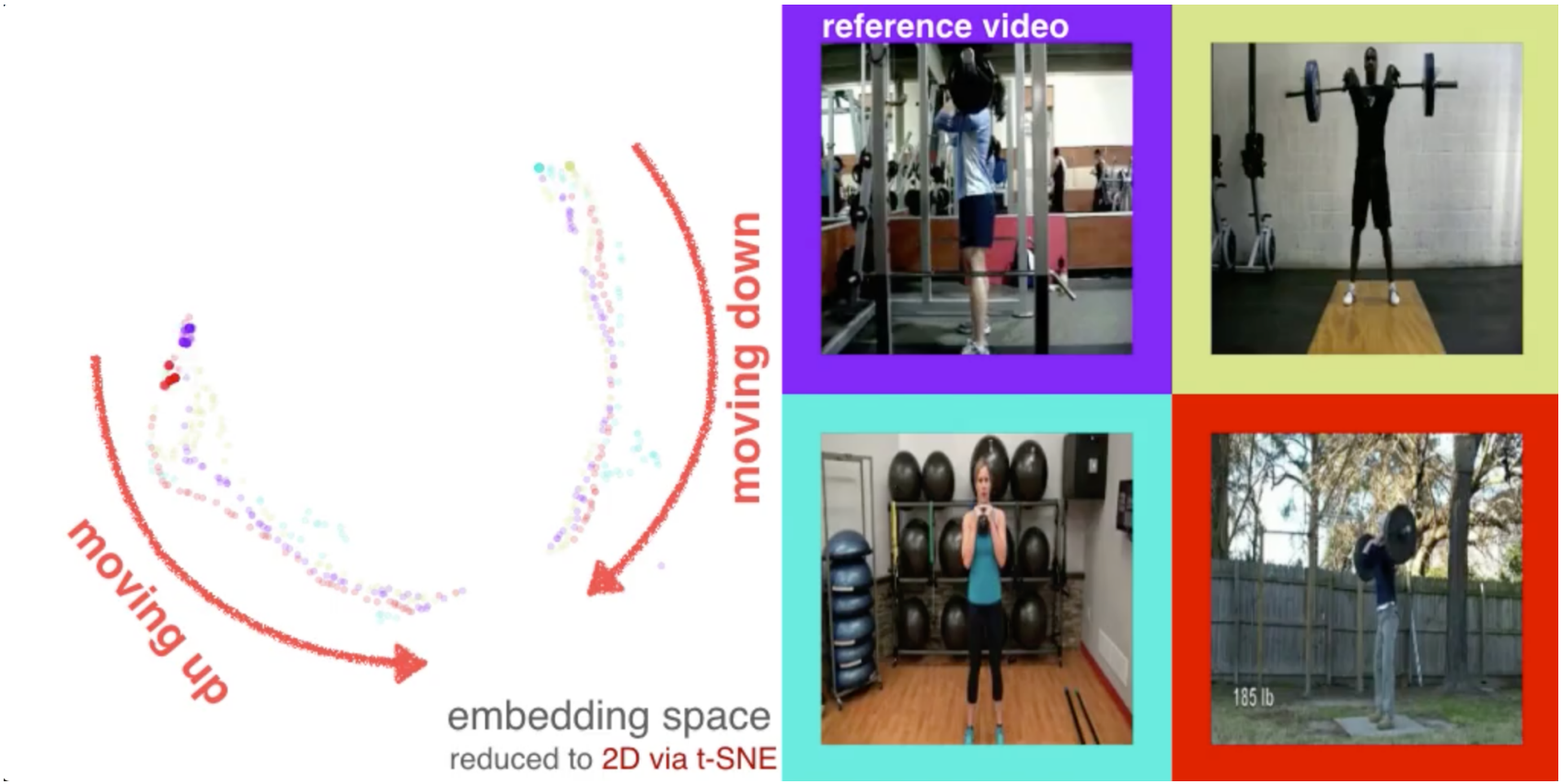
Key aspects:

- Soft nearest neighbor
- Differentiable cycle consistency

Image source: [Dwibedi et al. 2019]
Video source: https://debidatta.github.io/tcc_interactive_paper/index.html

Pretext learning

Temporal alignment between videos



Video source: https://debidatta.github.io/tcc_interactive_paper/index.html

Pretext learning

Temporal alignment between videos

Applications: Fine-grained retrieval

Image source: [Dwibedi et al. 2019]

Pretext learning

Temporal alignment between videos

Applications: Anomaly detection

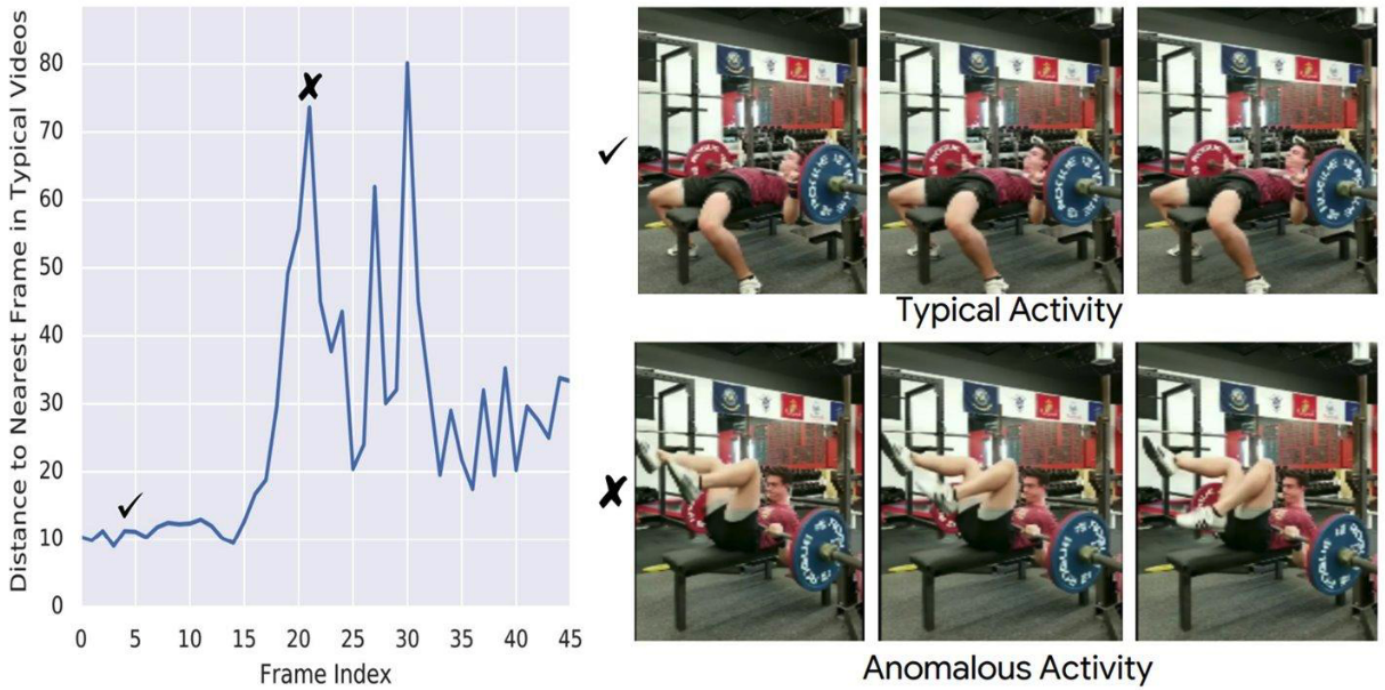


Image source: [Dwibedi et al. 2019]

Pretext learning

Temporal alignment between videos

Applications: Action phase classification (downstream task)

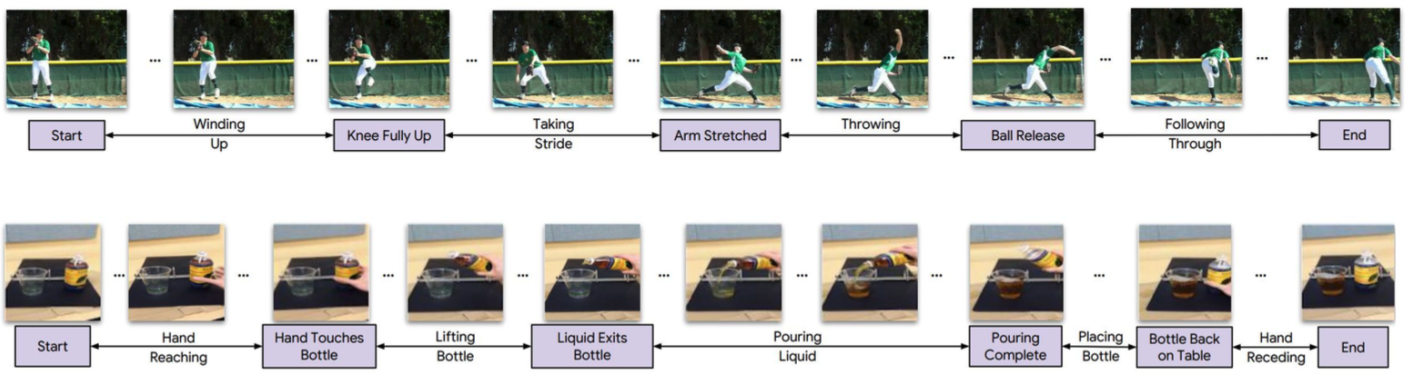


Image source: [Dwibedi et al. 2019]

Pretext learning

Temporal alignment between videos

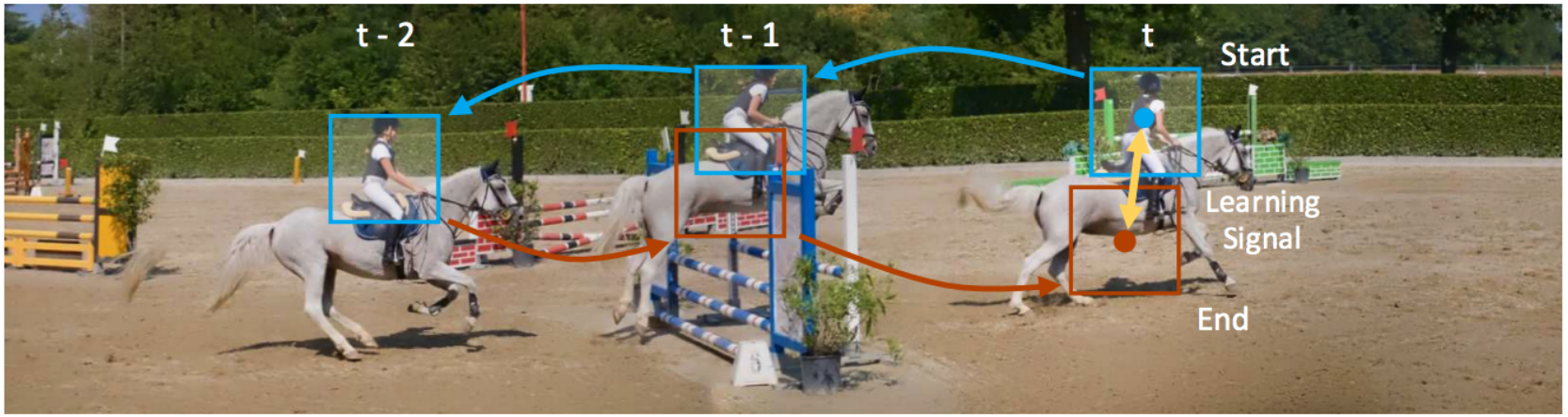
Other applications:

- Pace transfer
- Sound transfer
- Synchronization of multiple videos

Image source: [Dwibedi et al. 2019]

Pretext learning

Cycle-consistent tracking



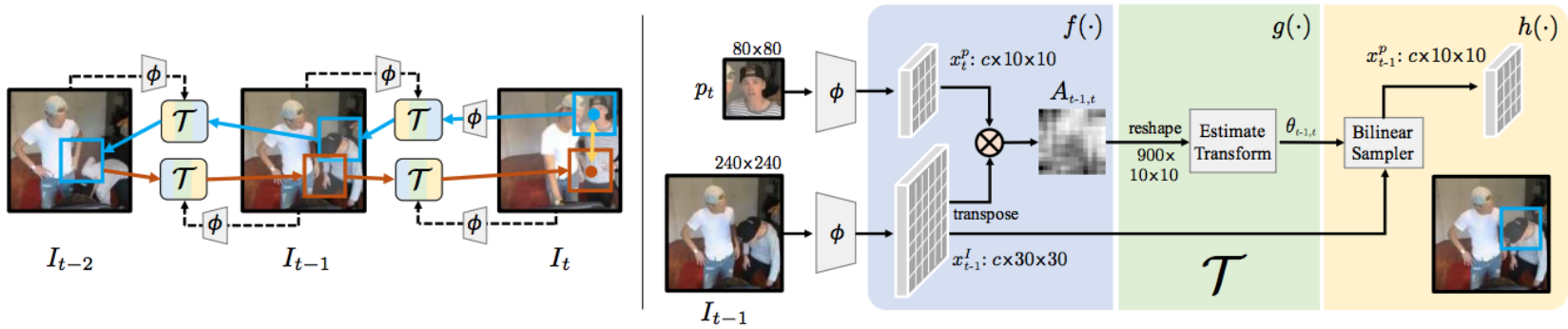
Idea: Use cycle-consistency in time as free supervisory signal for learning visual representations from scratch

X. Wang, A. Jabri, A.A. Efros. Learning Correspondence from the Cycle-consistency of Time. CVPR 2019.

Image source: [Wang et al. 2019]

Pretext learning

Cycle-consistent tracking



(a) Training ϕ by End-to-end Cycle-consistent Tracking

(b) Differentiable Tracking Operation $\mathcal{T}$

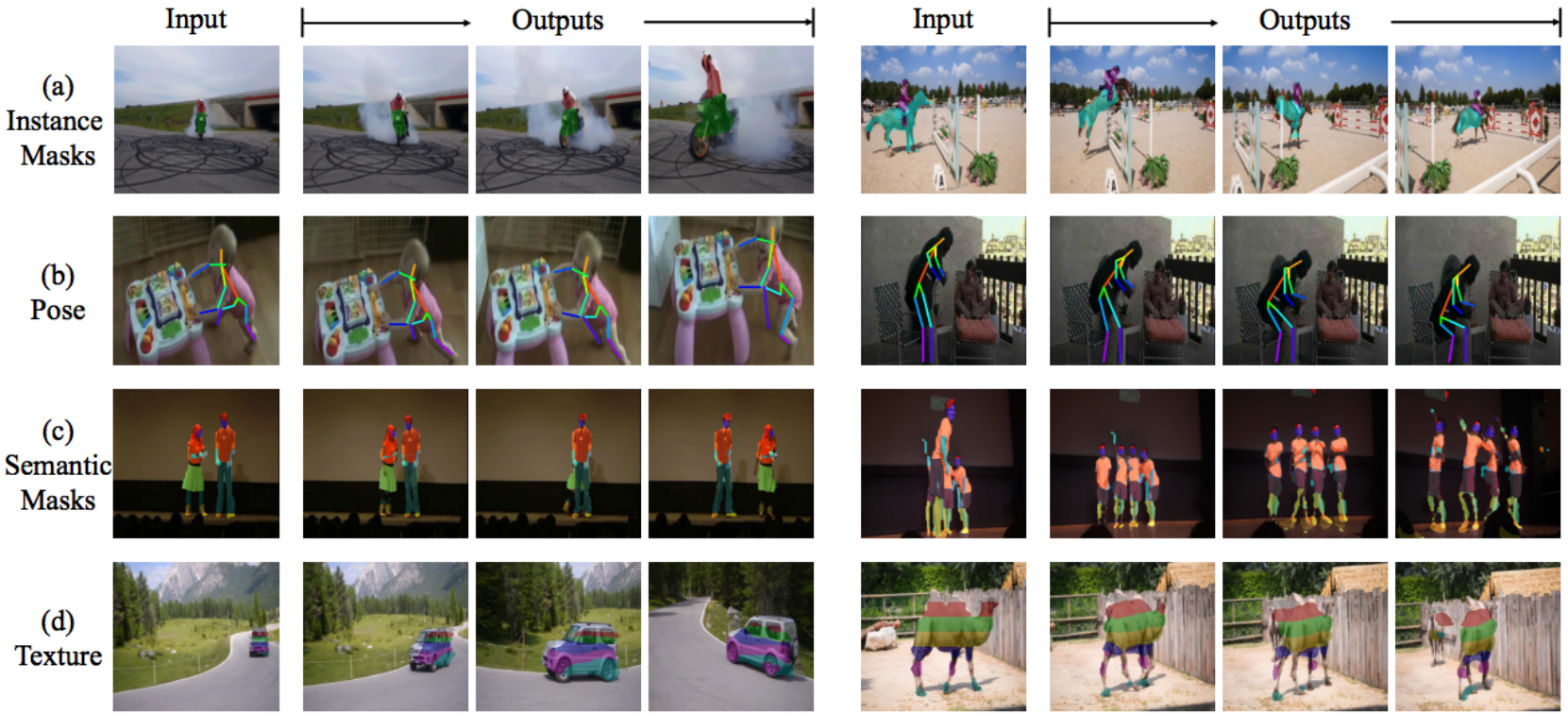
Note that:

- Only the initial patch p_t is explicitly encoded by ϕ; other patch features along the cycle are obtained by localizing image features.
- T is only used in training and is deliberately designed to be weak, so as to place the burden of representation on ϕ. At test time, the learned ϕ is used directly for computing correspondences.

Image source: [Wang et al. 2019]

Pretext learning

Cycle-consistent tracking



Applications: Tracking and propagation of labels (no fine-tuning)

Image source: [Wang et al. 2019]

Pretext learning

- Temporal order
  - Sequence verification
  - Sequence order
- Jigsaw puzzle
- Colorization
- Cycle consistency
  - Temporal alignment between videos
  - Cycle-consistent tracking
- Others: Appearance statistics, playback speed, ...

Outline

- Introduction to SSL
- SSL approaches in video
  - Pretext learning
  - **Generative learning**
  - Contrastive learning
  - Cross-modal agreement
- Multimodal and cross-modal approaches

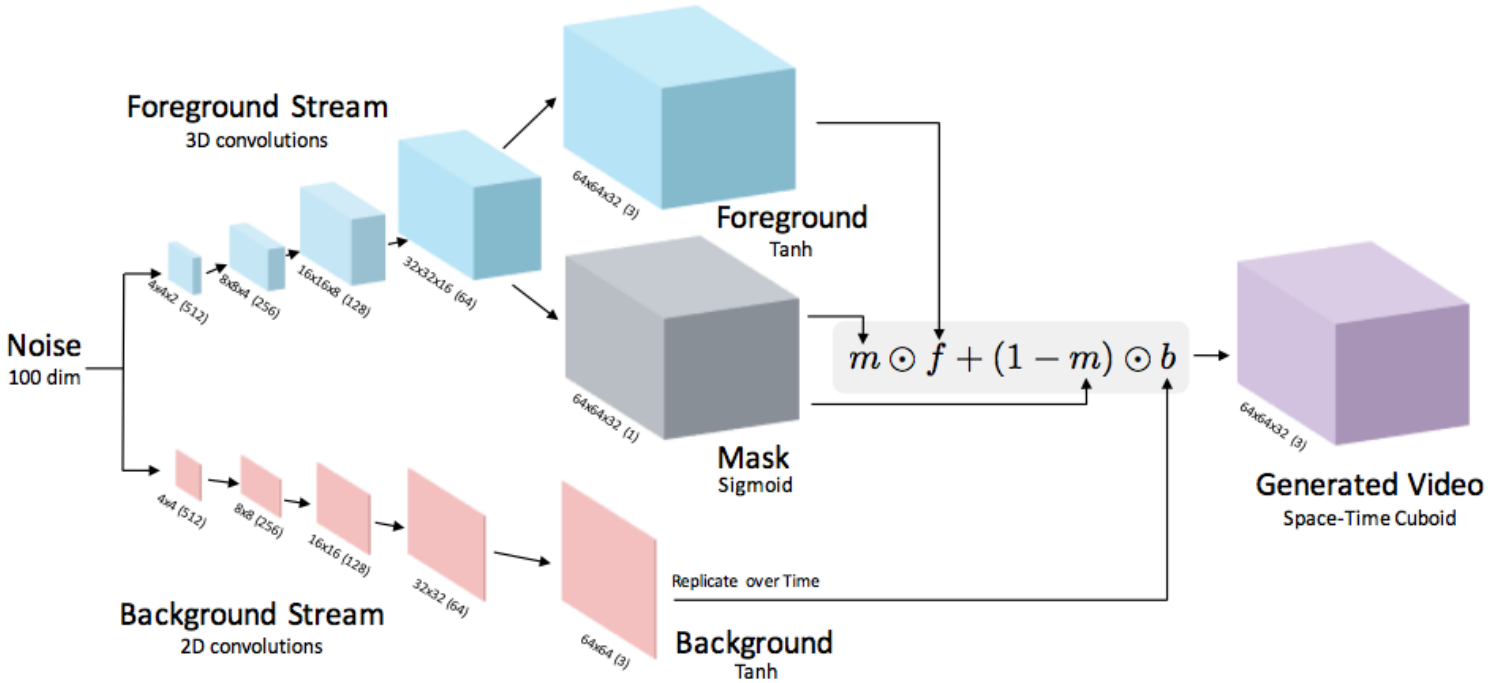
Generative learning

Different techniques [Schiappa et al. 2022]:

- Reconstruction of altered input/and or noise
- Generate future frames in a sequence (using previous frames)
- Generation from masked input (that can be patches in frames or entire frames)
- Generation from masked input of video, text and/or audio

Generative learning

Video GAN



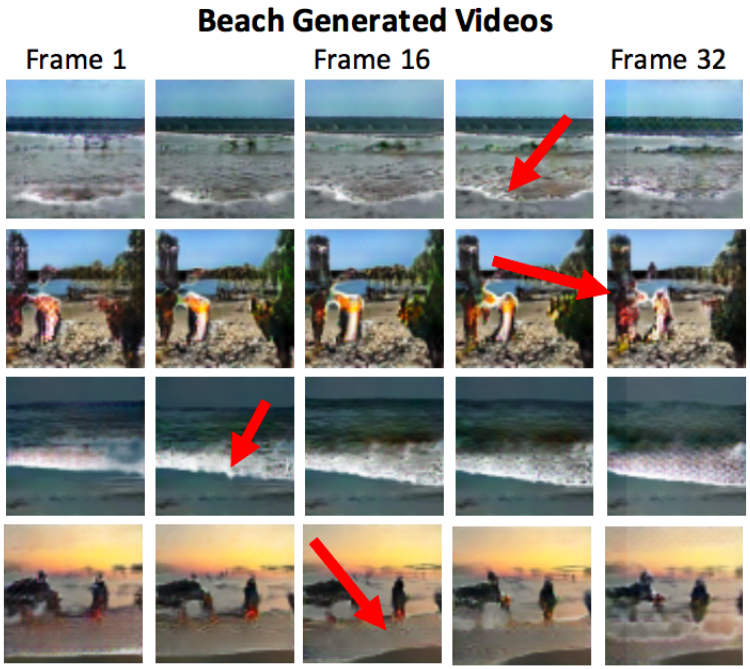
Scene dynamics are learned with an adversarial training.

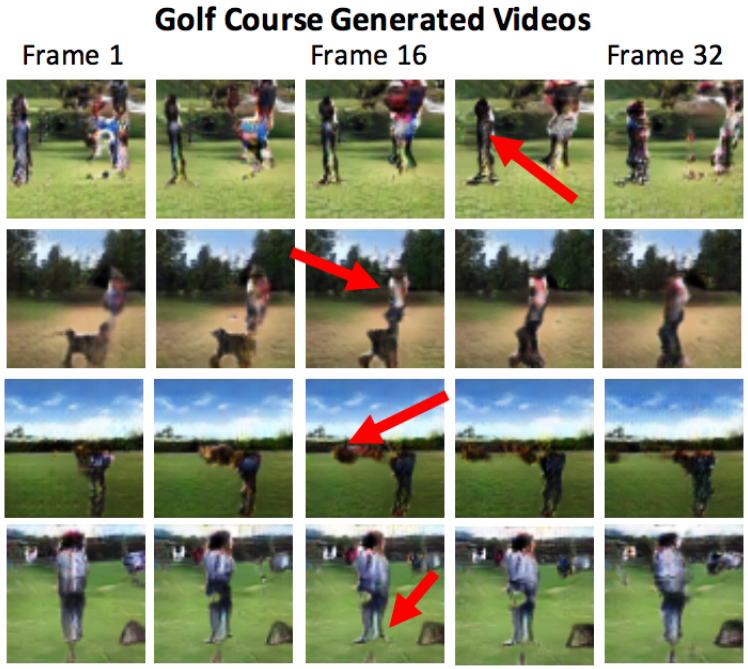
C. Vondrick, H. Pirsiavash, A. Torralba. Generating videos with scene dynamics. NeurIPs 2016.

Image source: [Vondrick et al. 2016]

Generative learning

Video GAN





The generated motion is often plausible

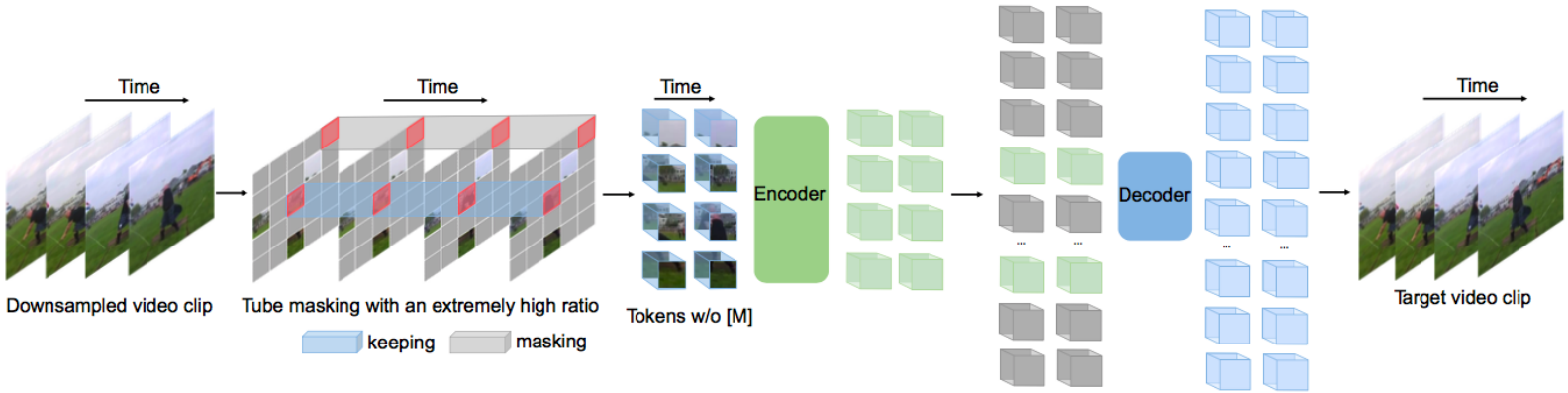
Downstream task: Action classification (by fine-tuning the discriminator)

C. Vondrick, H. Pirsiavash, A. Torralba. Generating videos with scene dynamics. NeurIPs 2016.

Image source: [Vondrick et al. 2016]

Generative learning

Video Masked Autoencoder



Proxy tast: Reconstruction

- Vanilla vision transformer (ViT) trained in a self-supervised way with masked autoencoder
- Tube masking with an extremely high ratio (due to temporal redundancy)

Downstream task: Action recognition

Z. Tong, Y. Song, J. Wang, L. Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. NeurIPS, 2022.

Image source: [Tong et al. 2022]

Outline

- Introduction to SSL
- SSL approaches in video
  - Pretext learning
  - Generative learning
  - **Contrastive learning**
  - Cross-modal agreement
- Multimodal and cross-modal approaches

Contrastive learning

Idea: Pull positive pairs and push negative pairs farther apart.

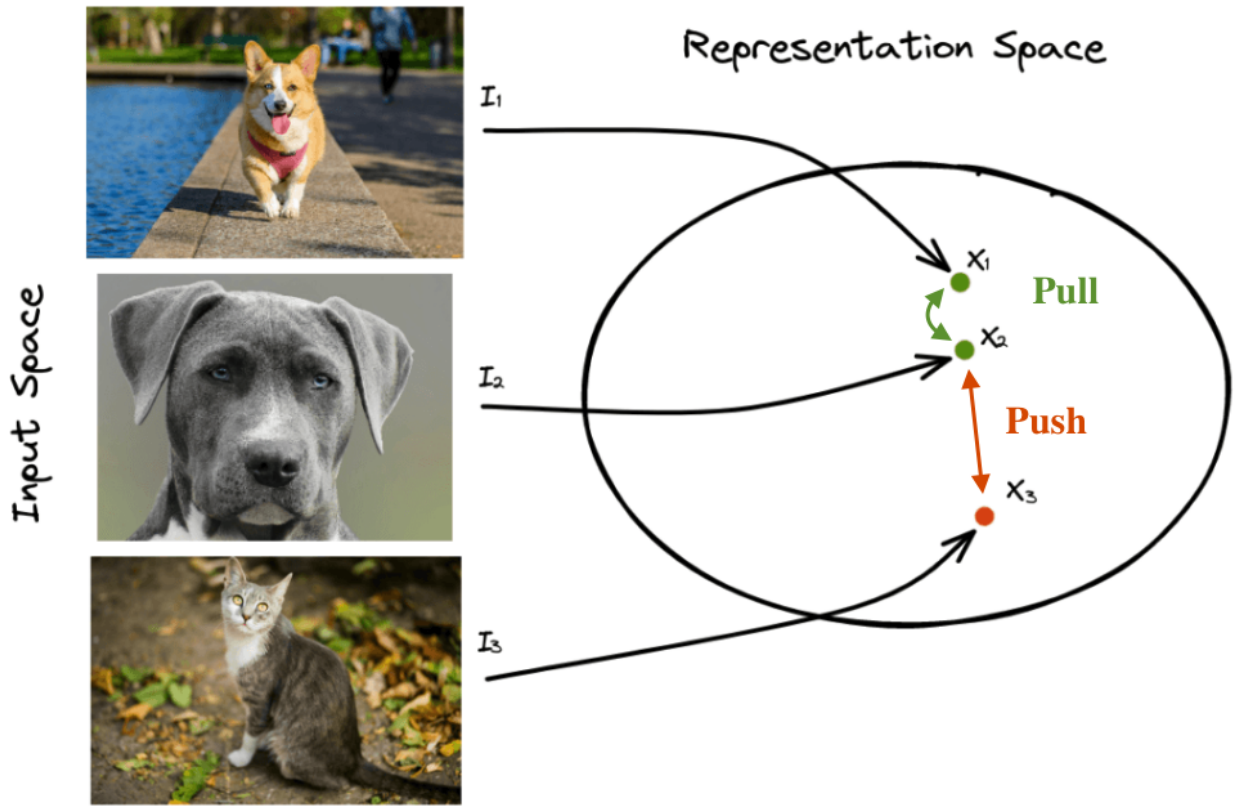
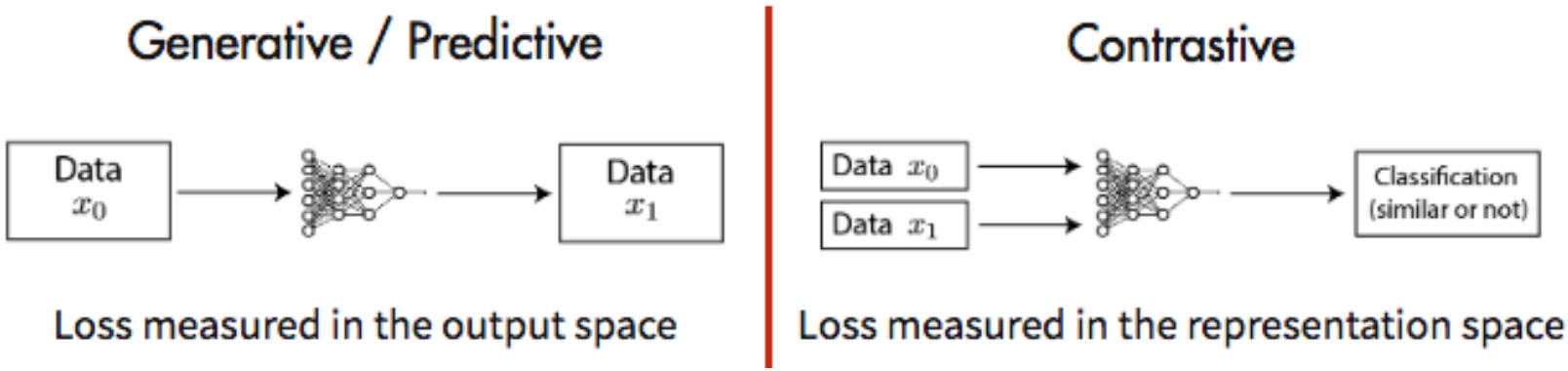


Image adapted from: https://www.baeldung.com/cs/contrastive-learning

Contrastive learning

Generative vs Contrastive methods



A loss in the representation space:

- captures more abstract latent factors (high-level features)
- better models correlations

Image source: https://ankeshanand.com/blog/2020/01/26/contrative-self-supervised-learning.html

Contrastive learning

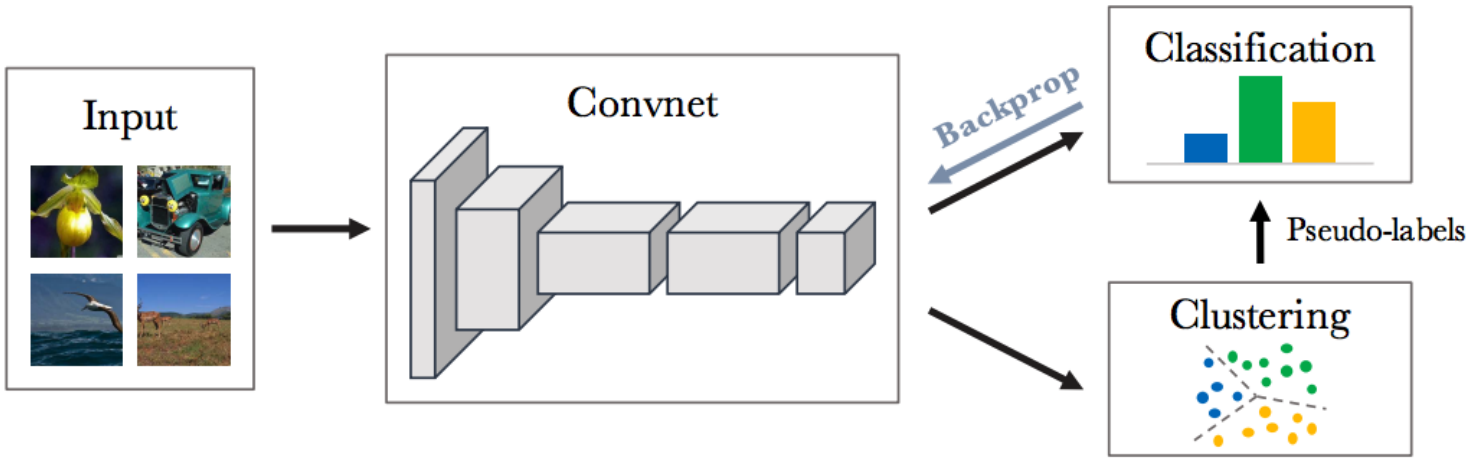
Idea: Pull positive pairs and push negative pairs farther apart.

Variations:

- How positive and negative samples are generated
  - Augmentation
  - Cross-modal agreement
  - Clustering (in the embedding space)
- Main objective
  - Classification (cross-entropy)
  - Discriminator
  - Variations of NCE loss

Contrastive learning

Deep clustering



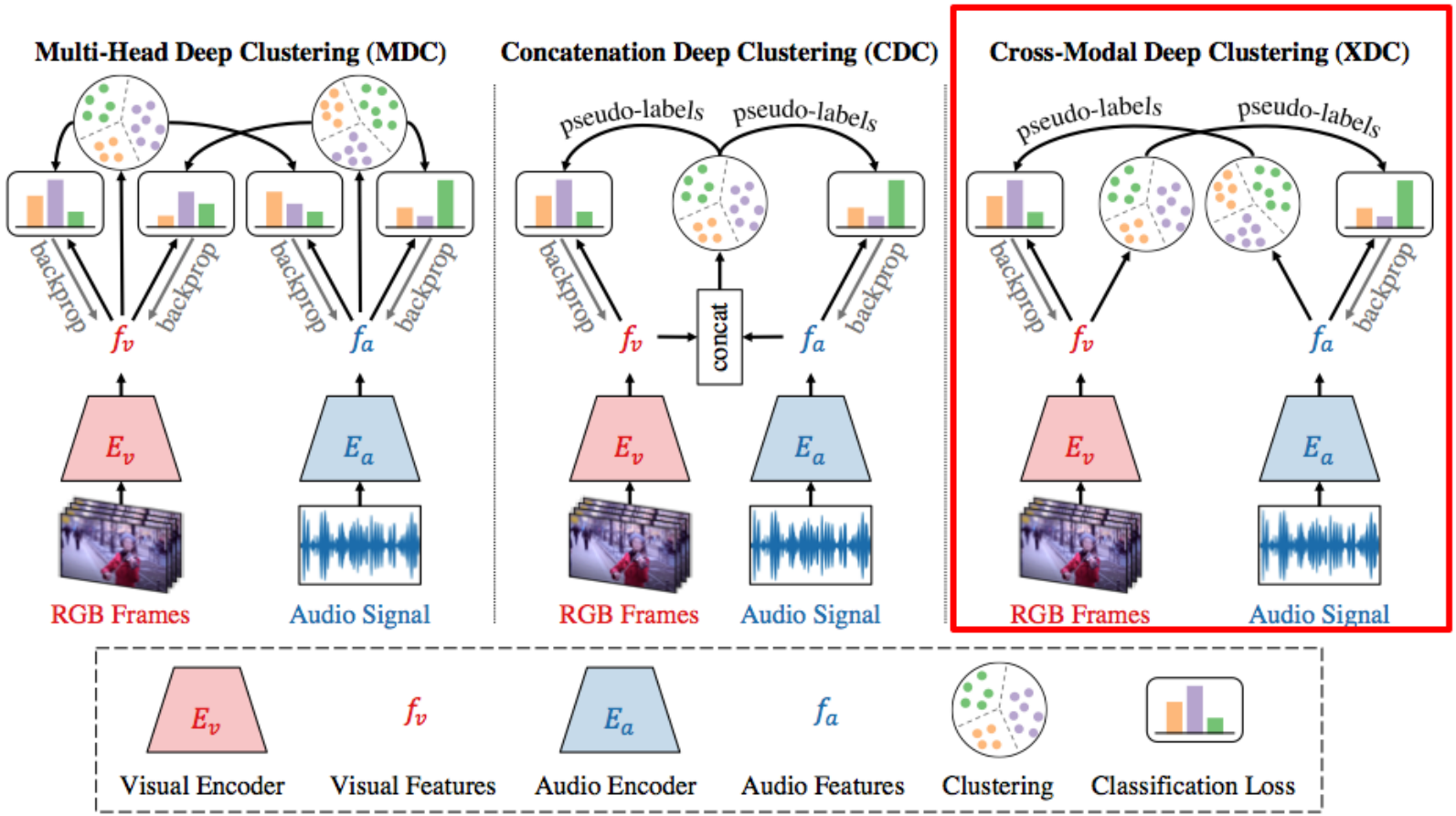
Downstream tasks: Image classification, object detection and semantic segmentation

M. Caron, P. Bojanowski, A. Joulin, M. Douze. Deep clustering for unsupervised learning of visual features. ECCV 2018.

Image source: [Caron et al. 2018]

Contrastive learning

Cross-modal deep clustering



Downstream tasks: Action recognition and audio classification

H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, D. Tran. Self-supervised learning by cross-modal audio-video clustering. Generating videos with scene dynamics. NeurIPs, 2020.

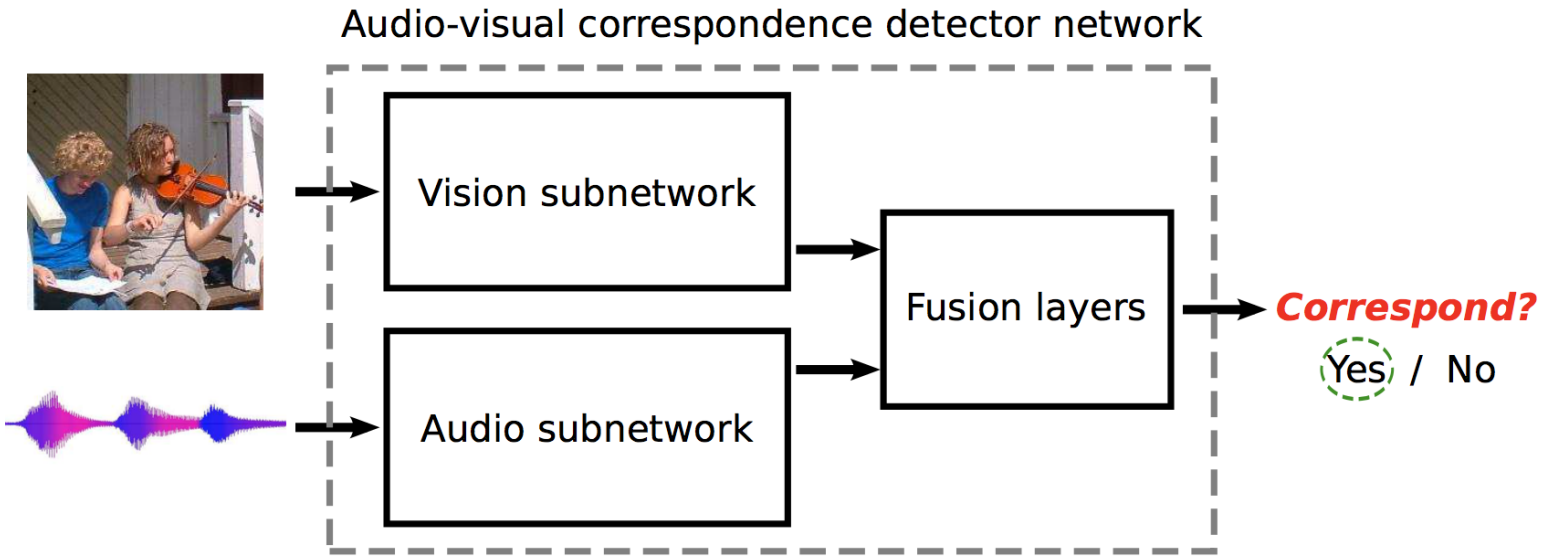
Image source: [Alwassel et al. 2020]

Outline

- Introduction to SSL
- SSL approaches in video
  - Pretext learning
  - Generative learning
  - Contrastive learning
  - **Cross-modal agreement**
- Multimodal and cross-modal approaches

Cross-modal agreement

Audio-visual correspondence



Downstream tasks: Sound classification, image classification

R. Arandjelovic, A. Zisserman. Look, listen and learn. ICCV 2017.

Image source: [Arandjelovic and Zisserman 2017]

Cross-modal agreement

Audio-visual correspondence

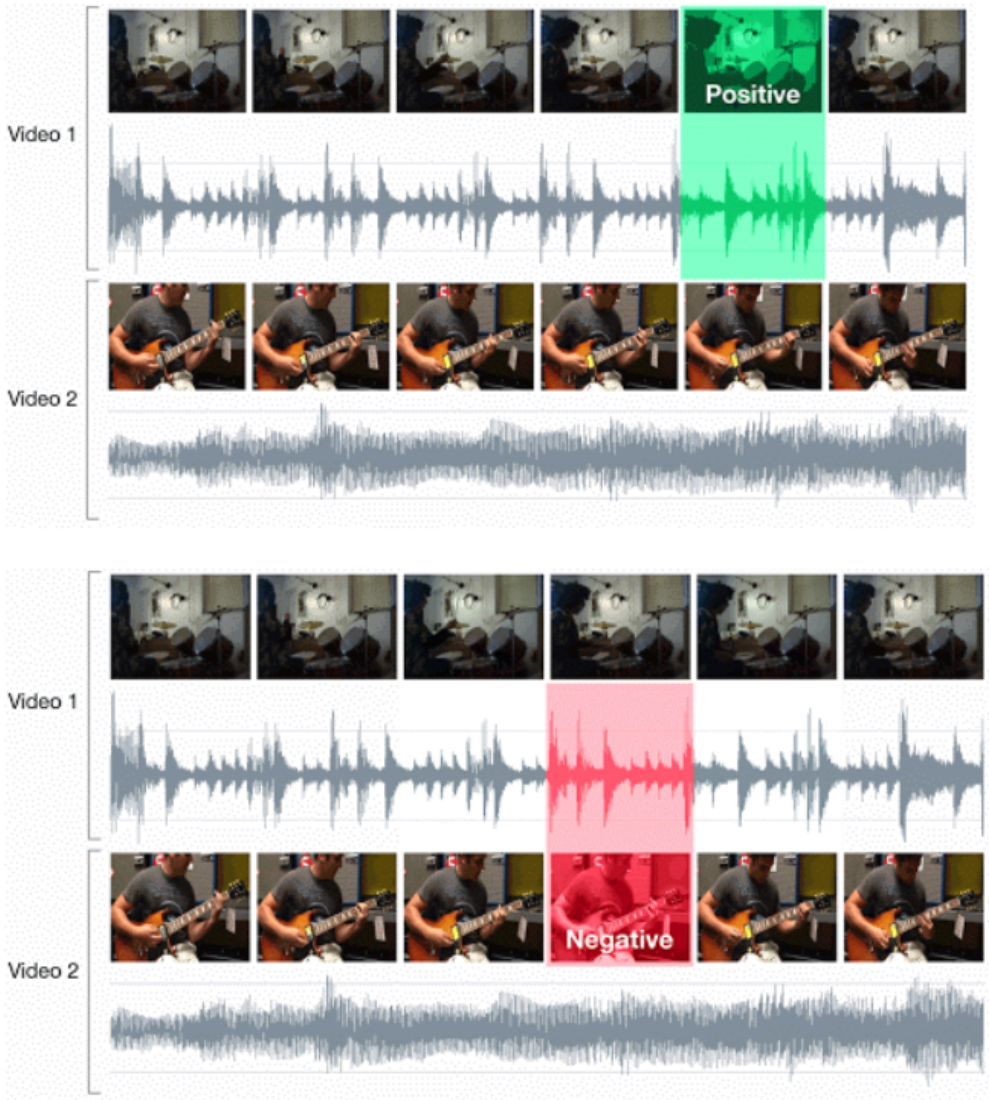
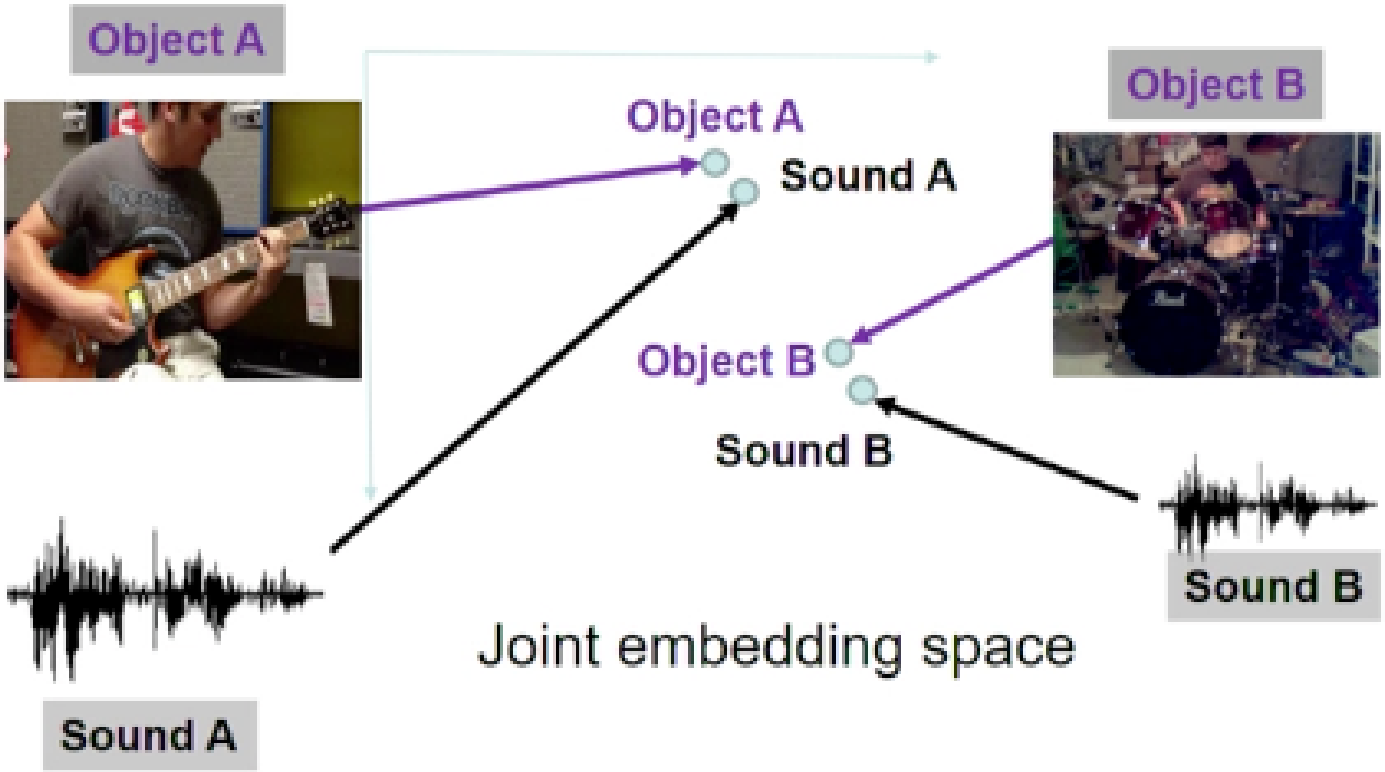


Image source: [Arandjelovic and Zisserman 2018]

Cross-modal agreement

Audio-visual correspondence



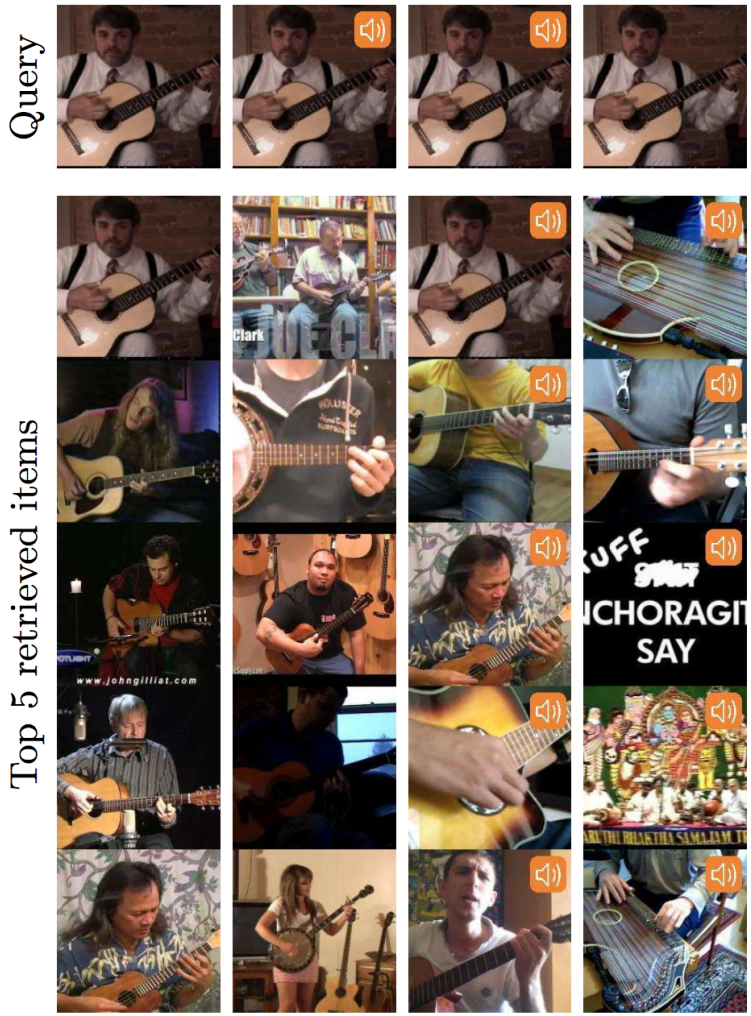
Downstream tasks: Sound classification, image classification
Applications: Intra-modal and crosss-modal retrieval, sound localization.

R. Arandjelovic, A. Zisserman. Objects that sound. ECCV 2018.

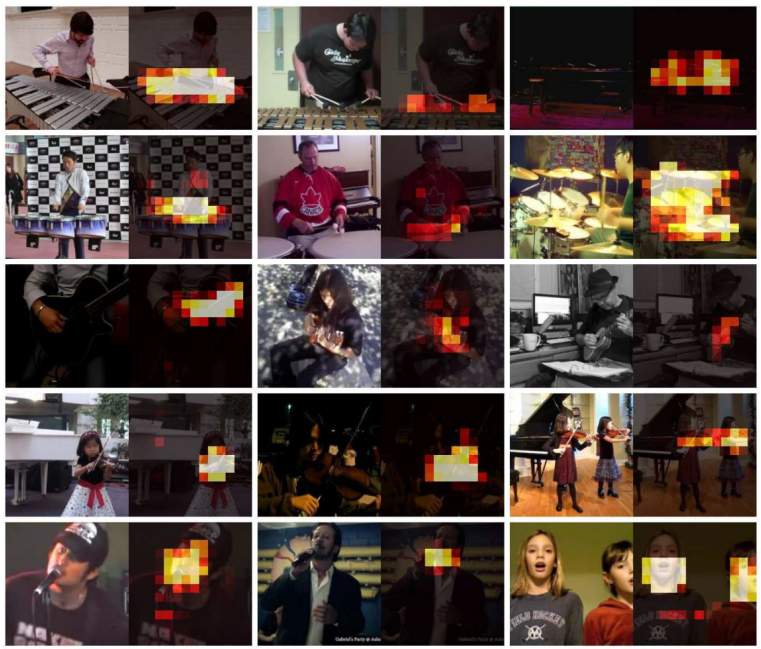
Image source: [Arandjelovic and Zisserman 2018]

Cross-modal agreement

Audio-visual correspondence



Intra-modal and crosss-modal retrieval

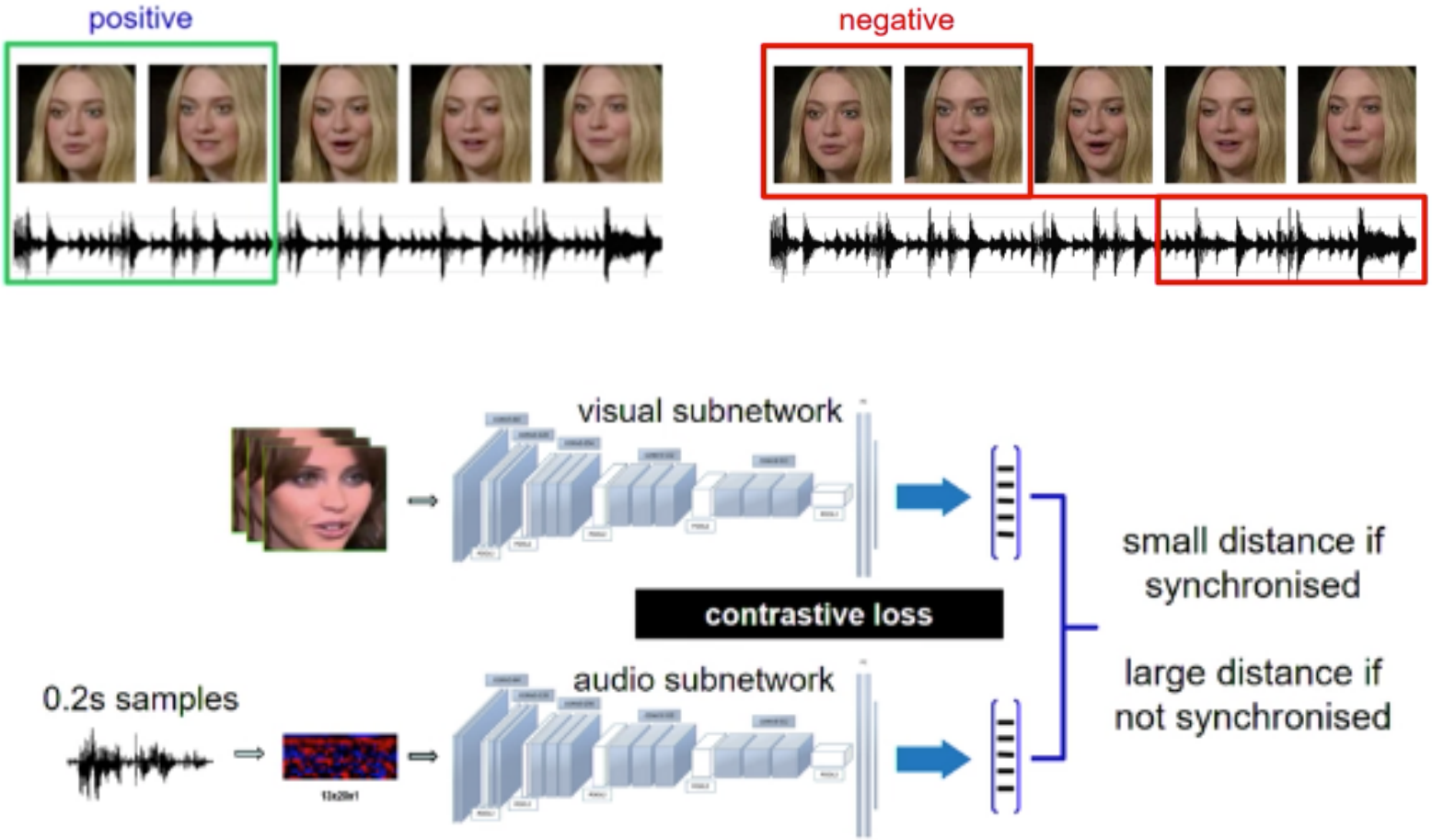


sound localization

Image source: [Arandjelovic and Zisserman 2018]

Cross-modal agreement

Audio-visual synchronization

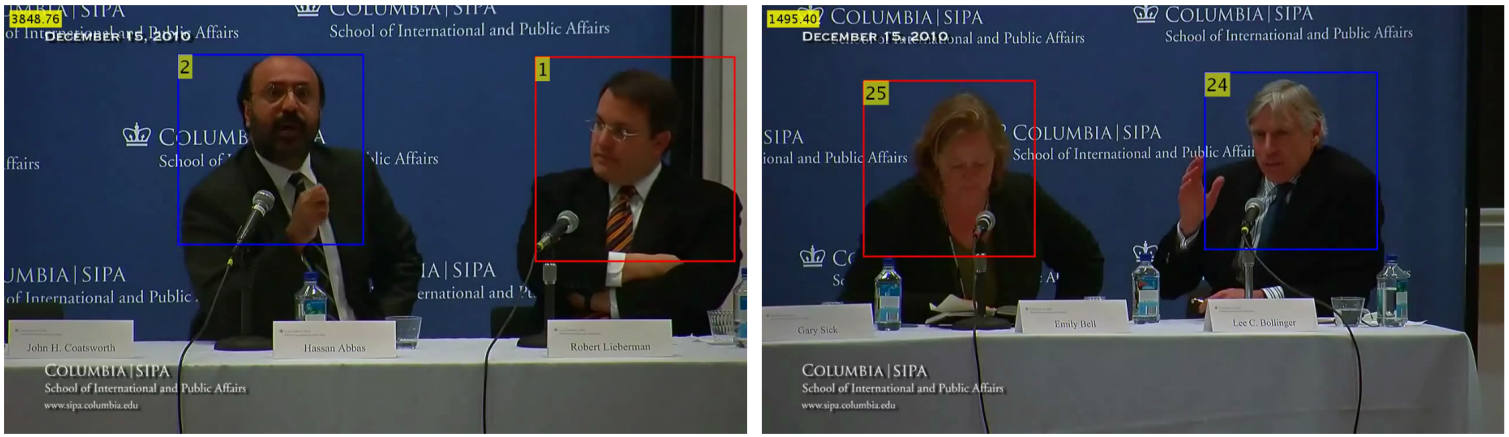


J.S. Chung, A. Zisserman, Out of time: automated lip sync in the wild. ACCVw 2016.

Image source: [Chung and Zisserman 2016]

Cross-modal agreement

Audio-visual synchronization



Speaker Non-speaker

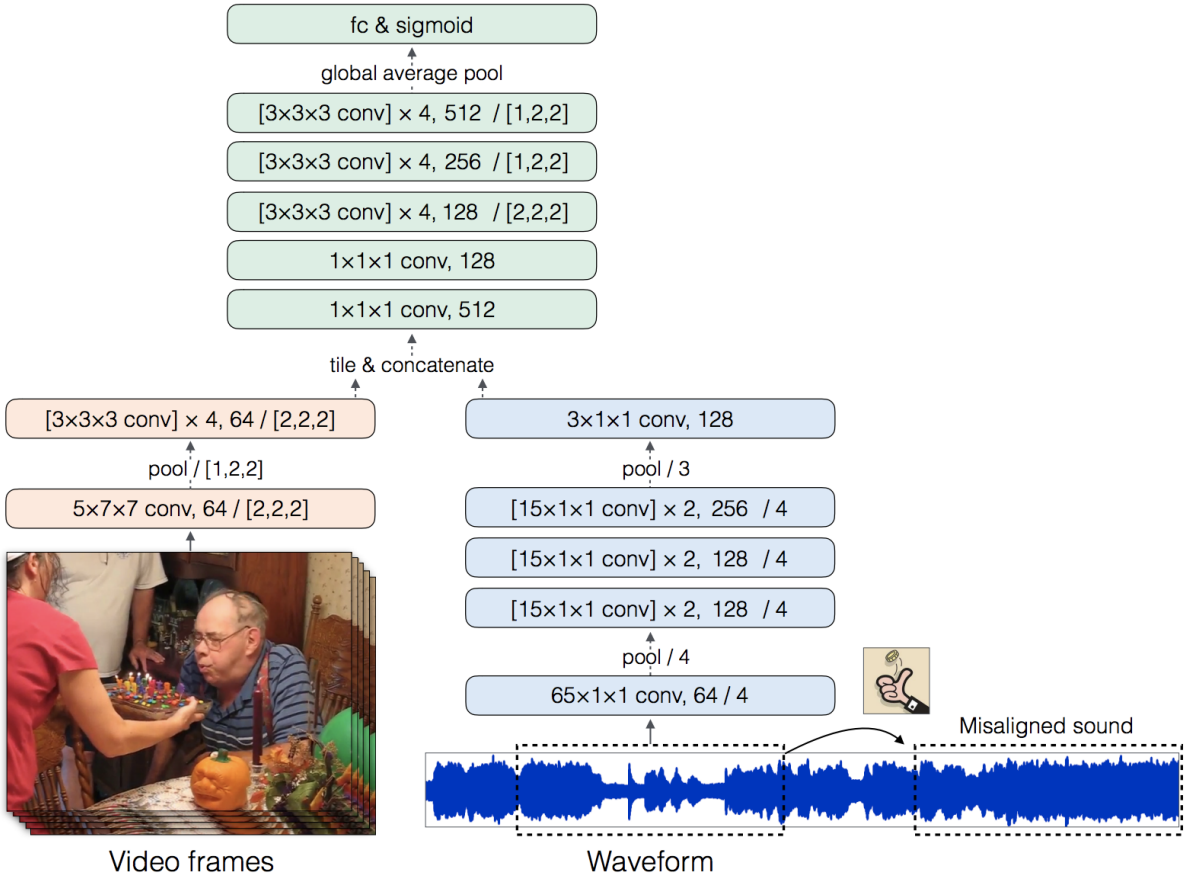
Application: Active speaker detection, Lip reading

J.S. Chung, A. Zisserman, Out of time: automated lip sync in the wild. ACCVw 2016.

Image source: [Chung and Zisserman 2016]

Cross-modal agreement

Audio-visual synchronization



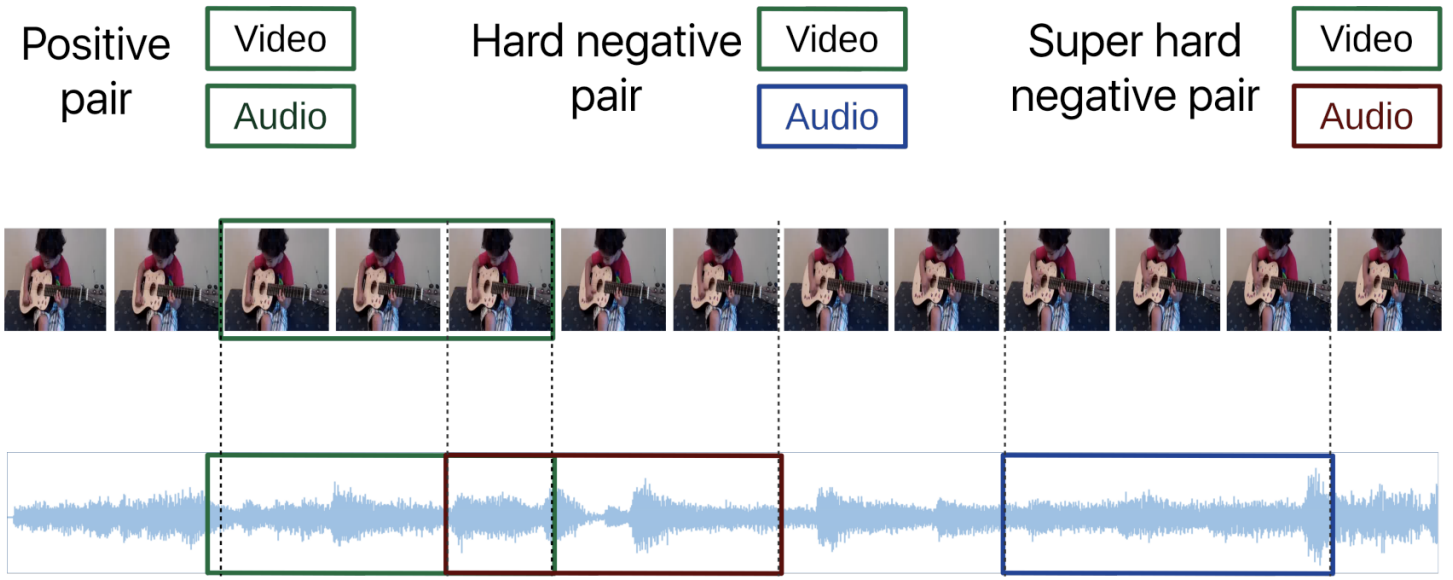
Learned visual and acoustic features useful for: Sound source localization, audio-visual action recognition, on/offscreen audio source separation.

A. Owens, A. A. Efros. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. ECCV 2018.

Image source: [Owens and Efros 2018]

Cross-modal agreement

Audio-visual synchronization



Downstream tasks: Action recognition, audio classification

B. Korbar, D. Tran, L. Torresani. Cooperative learning of audio and video models from self-supervised synchronization. NeurIPS 2018.

Image source: [Korbar et al. 2018]

Cross-modal agreement

Learning from misaligned and noisy narrations

Challenges:

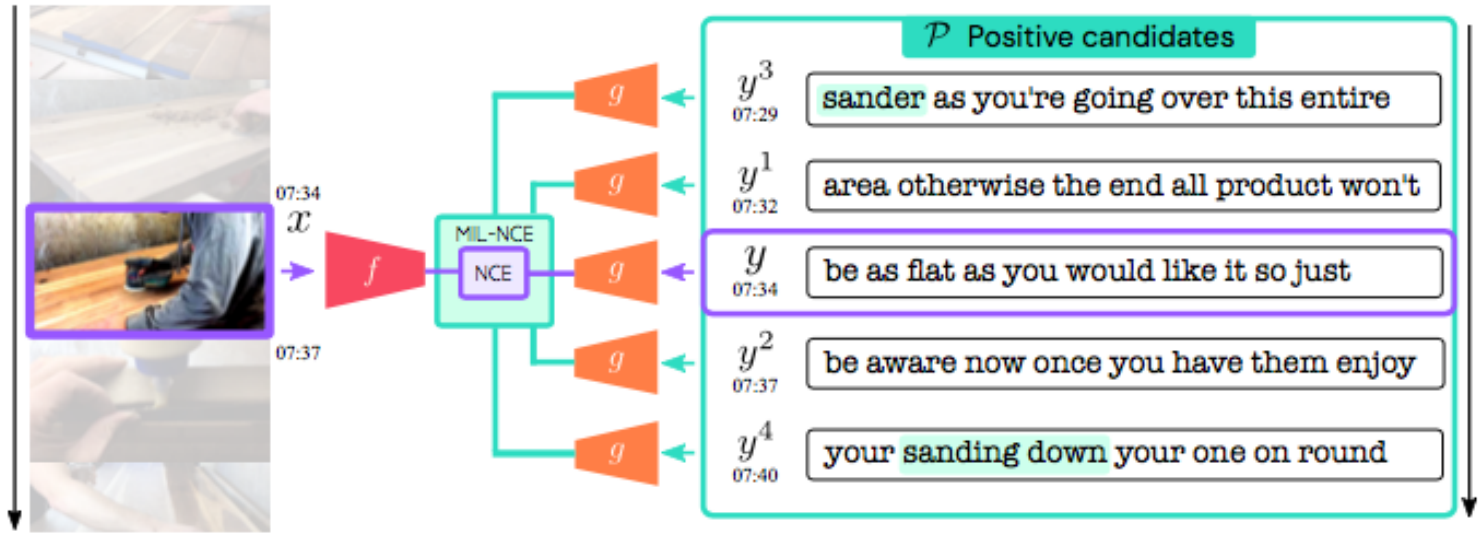
- Weak alignment (narration of an action before or after demonstrating it)
- Certain information only present in one modality
- Irrelevant information in the speech (e.g. jokes or credits)
- Erroneous speech recognition

A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, A. Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. CVPR 2020

Image source: [Miech et al. 2020]

Cross-modal agreement

Learning from misaligned and noisy narrations



$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)}}{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)} + \sum_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}} \right)$$

MIL-NCE

$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{e^{f(x_i)^\top g(y_i)}}{e^{f(x_i)^\top g(y_i)} + \sum_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}} \right)$$

InfoNCE

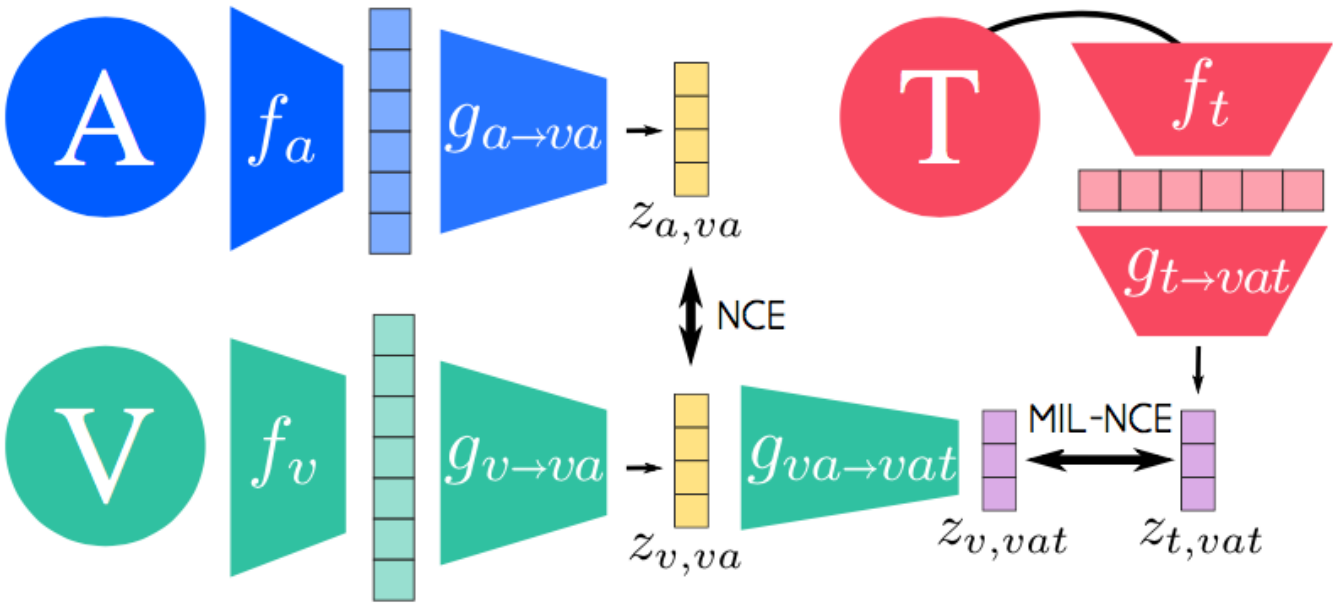
Downstream tasks: Action recognition, text-to-video retrieval, action localization and action segmentation.

A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, A. Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. CVPR 2020

Image source: [Miech et al. 2020]

Cross-modal agreement

Learning from audio, video and text



Downstream tasks: Action recognition, sound classification, text-to-video retrieval, image classifications.

J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, A. Zisserman. Self-Supervised MultiModal Versatile Networks. NeurIPS 2020.

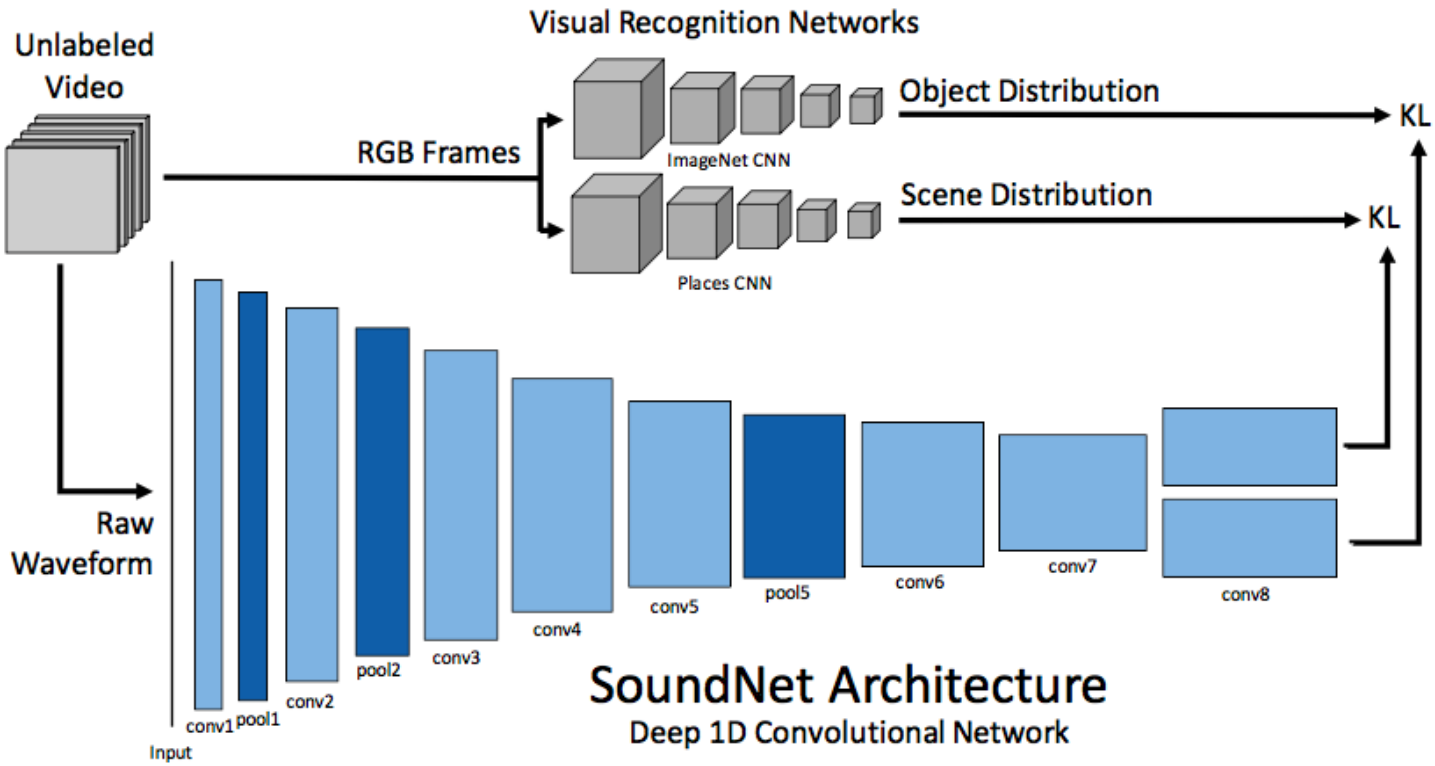
Image source: [Alayrac et al. 2020]

Outline

- Introduction to SSL
- SSL approaches in video
  - Pretext learning
  - Generative learning
  - Contrastive learning
  - Cross-modal agreement
- **Multimodal and cross-modal approaches**

Multimodal and cross-modal approaches

Cross-modal distillation

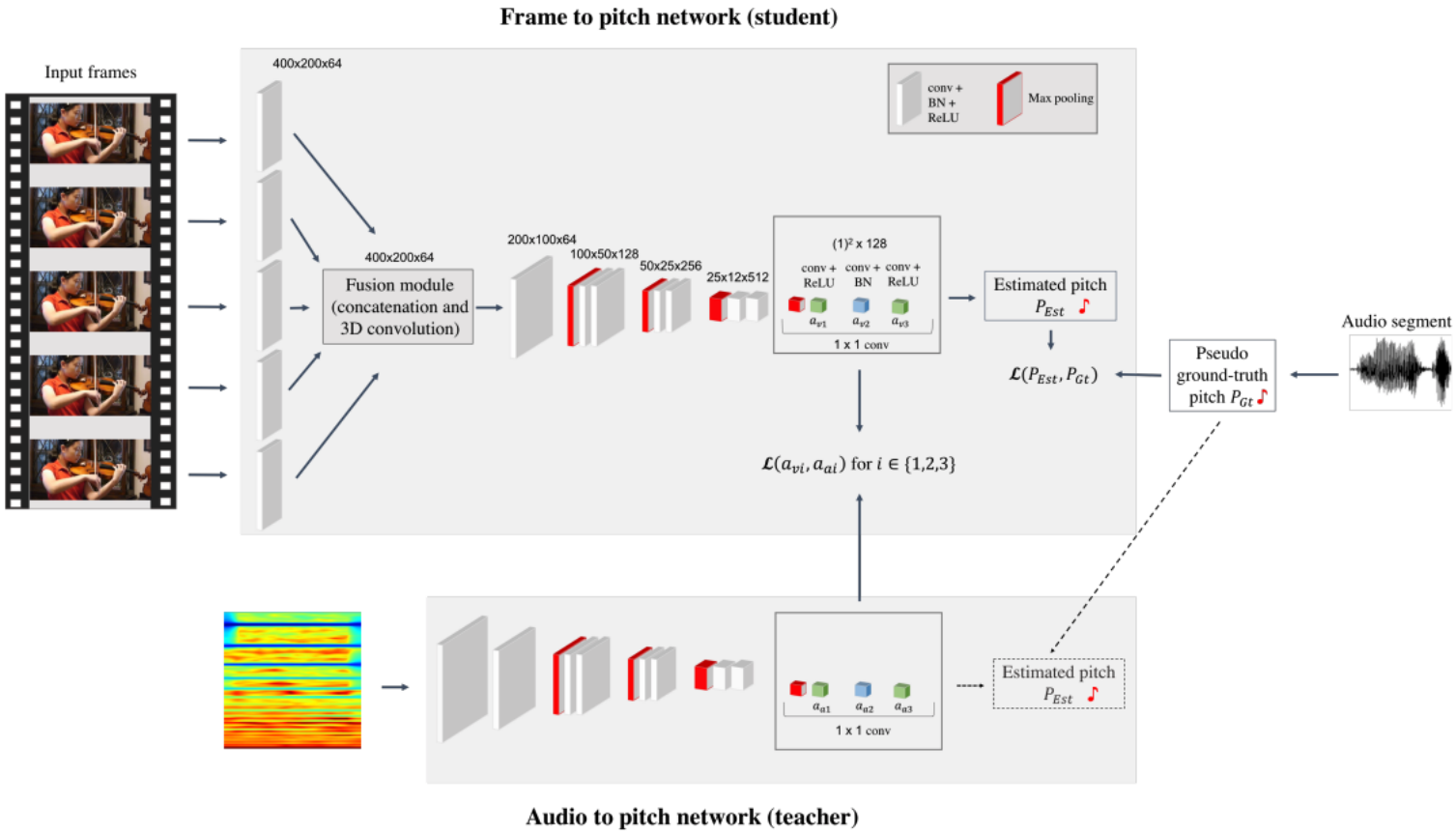


Y. Aytar, C. Vondrick, A. Torralba. SoundNet: Learning Sound Representations from Unlabeled Video. NIPS 2016

Image source: [Aytar et al. 2016]

Multimodal and cross-modal approaches

Cross-modal distillation

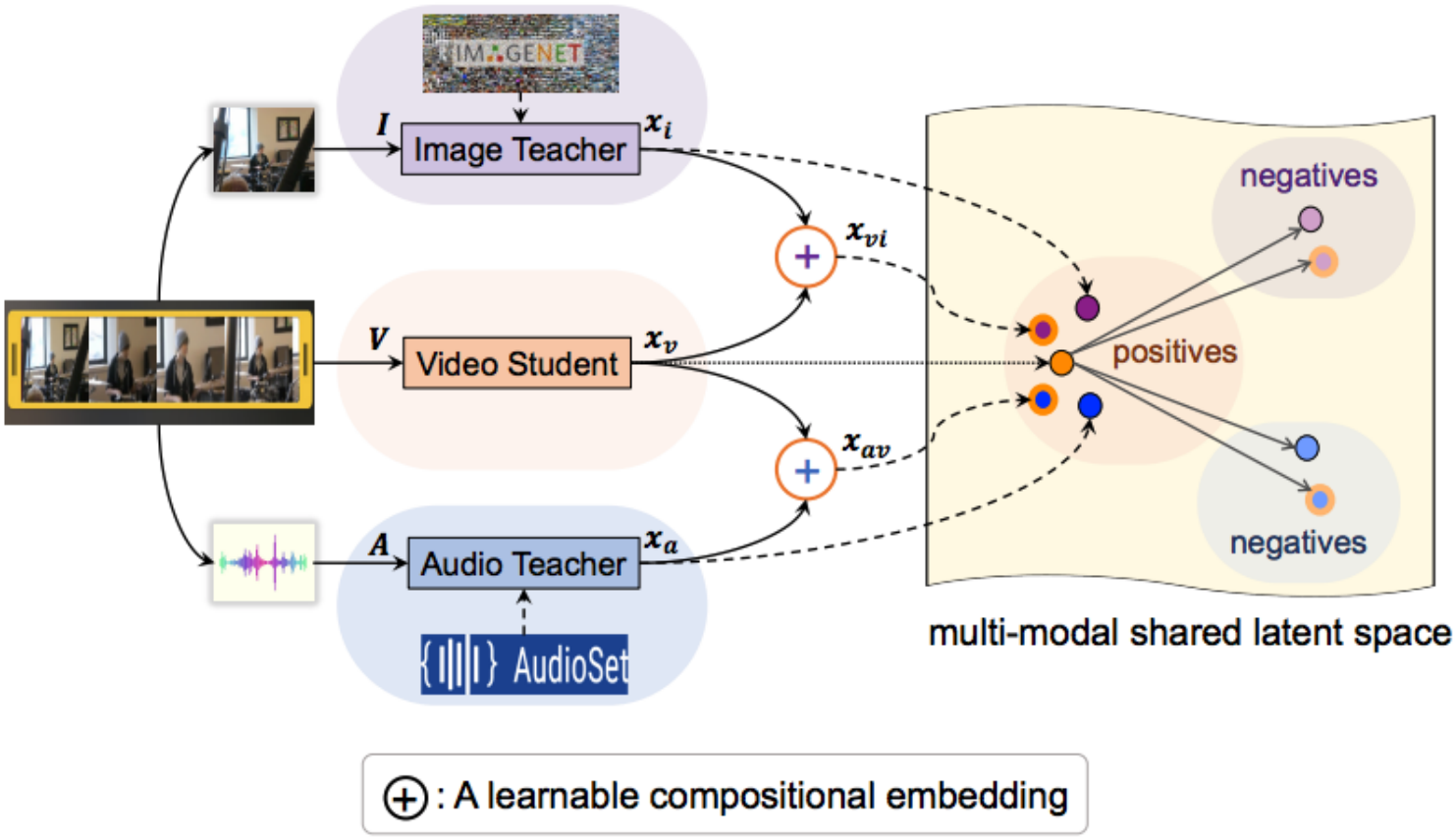


A. S. Koepke, O. Wiles, A. Zisserman. Visual pitch estimation. SMC, 2019

Image source: [Chen et al. 2021]

Multimodal and cross-modal approaches

Cross-modal distillation



Y. Chen, Y. Xian, A. S. Koepke, Y. Shan, Z. Akata. Distilling Audio-Visual Knowledge by Compositional Contrastive Learning, CVPR 2021

Image source: [Koepke et al. 2019]

Multimodal and cross-modal approaches

Video to audio generation

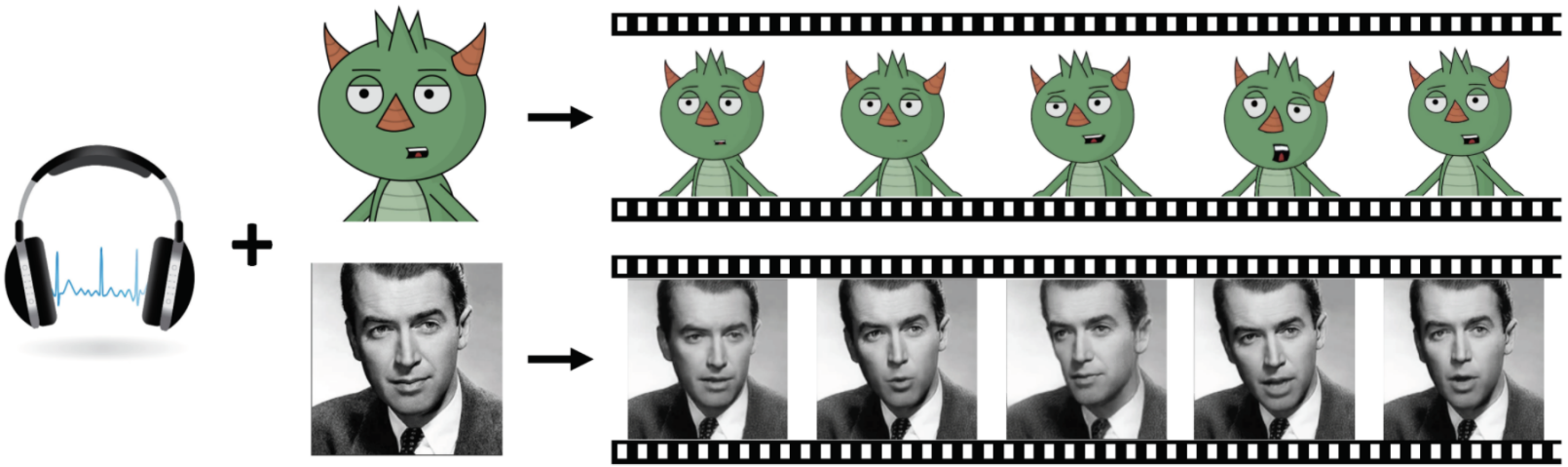
Video: https://andrewowens.com/vis/

A. Owens, P. Isola, J. McDermott, A. Torralba, E.H. Adelson, W.T. Freeman. Visually indicated sounds. CVPR 2016.

Image source: [Owens et al. 2016]

Multimodal and cross-modal approaches

Audio + image to video generation



Input: audio and single portrait image

Output: talking head animation

Video: https://people.umass.edu/~yangzhou/MakeItTalk/

Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, D. Li. MakeItTalk: Speaker-Aware Talking-Head Animation, SIGGRAPH ASIA 2020.

Image source: [Zhou et al. 2020]

Multimodal and cross-modal approaches

Sound source separation

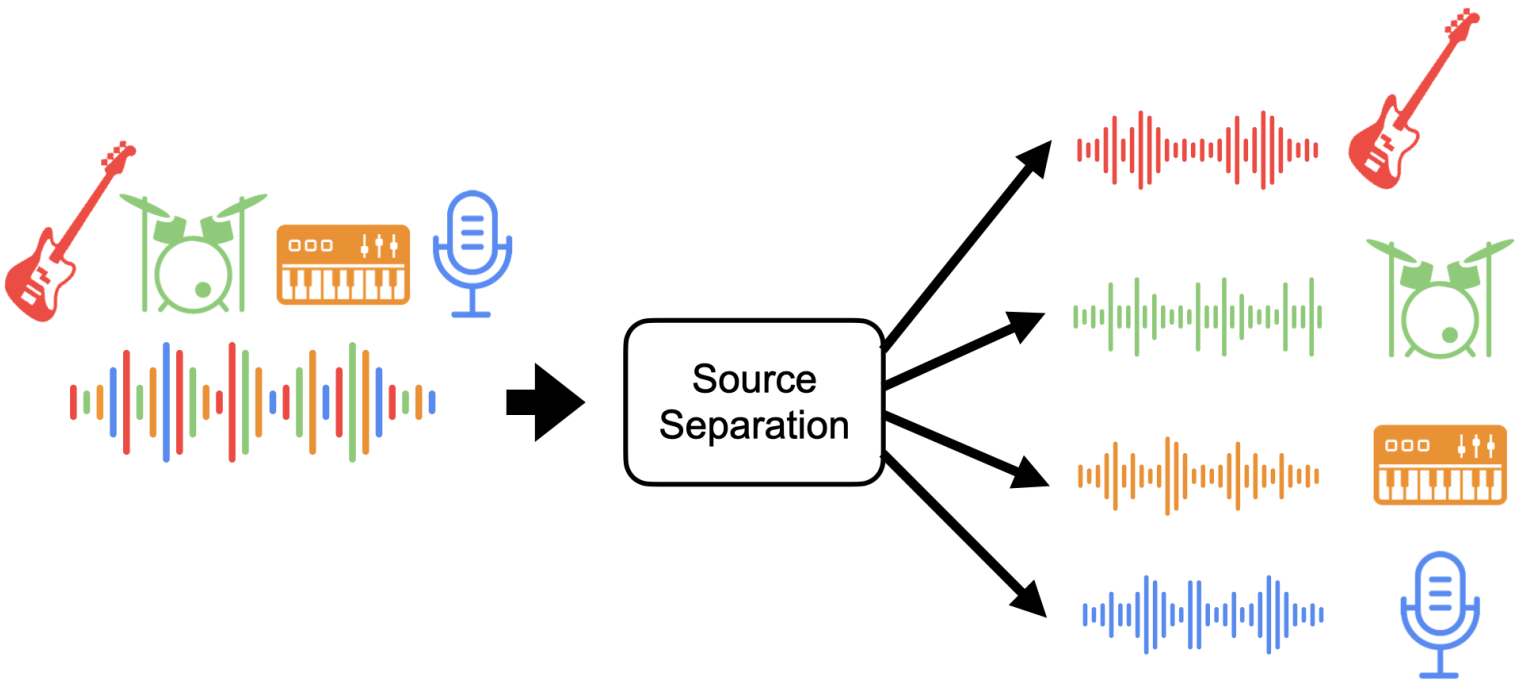
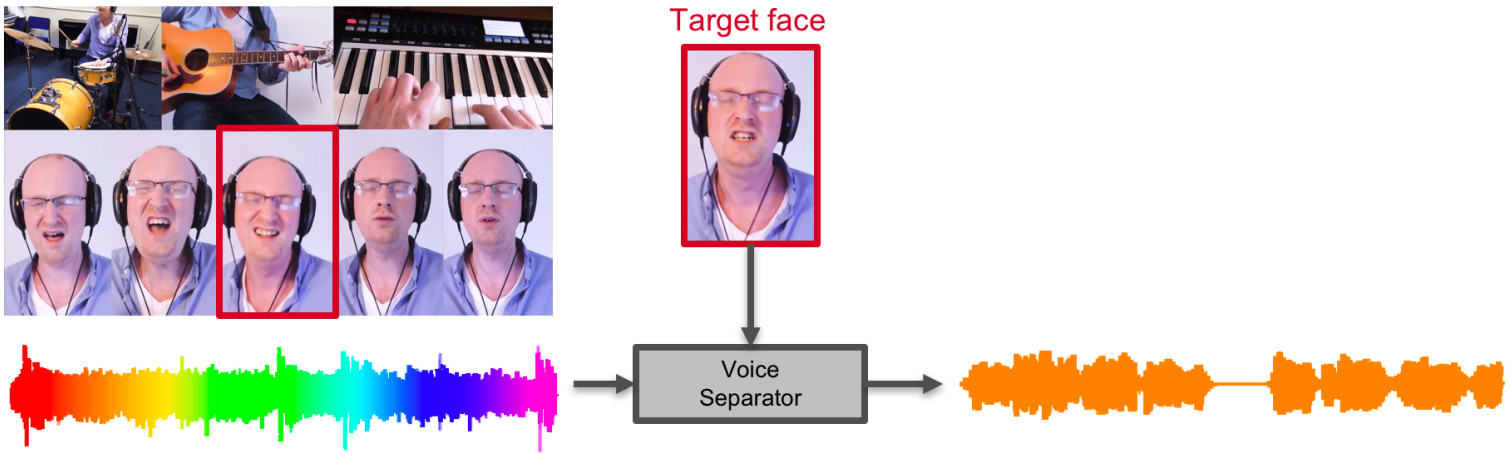


Image source: https://source-separation.github.io/tutorial/landing.html SSSS

Multimodal and cross-modal approaches

Sound source separation



First audio-visual speech separation methods (concurrent):

T. Afouras, J. S. Chung and A. Zisserman. The Conversation: Deep Audio-Visual Speech Enhancement. Interspeech 2018.
A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, M. Rubinstein. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. SIGGRAPH 2018.
A. Owens, A. A. Efros. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. ECCV 2018.

Recent methods:
https://vision.cs.utexas.edu/projects/VisualVoice/
https://ipcv.github.io/VoViT/

Summary

We have seen:

- Introduction to SSL
- SSL approaches in video
  - Pretext learning
  - Generative learning
  - Contrastive learning
  - Cross-modal agreement
- Multimodal and cross-modal approaches