Self-supervised learning of visual representations from video and natural language

Josef Šivic

CZECH TECHNICAL
UNIVERSITY
IN PRAGUE

Visual recognition

… extracting information from images

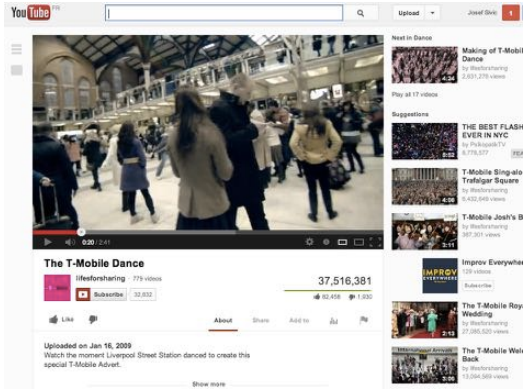
What human brain sees

What computer sees:

array of pixel intensities

Towards collective visual memory

Archives of visual information



Internet videos

10,000+ **TV channels**

Historical imagery

Cameras around us

2M+ **surveillance cameras**

Car cameras

Personal cameras

Record over time visual experiences of many people at different places into an emerging collective visual memory

Motivation

What if we could automatically learn from this visual data?

Learn from people to sequences of manipulation actions to achieve a certain task

"How to" instructional videos

Potential impact: machines that learn from collective visual memory for robotics

Motivation

What if we could automatically learn from this visual data?

Machines that autonomously learn to perceive, reason and act.

To operate in dangerous environments
[Darpa robot challenge 2015]

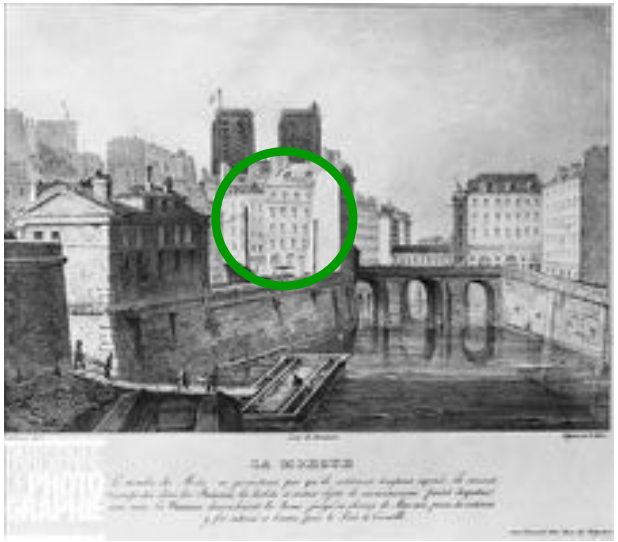
To assist people
[Microsoft HoloLens 2015]

Potential impact: machines that learn from collective visual memory for robotics

Motivation

What if we could automatically learn from this visual data?

Evolution of a particular place over time



1830

1852

1900

Potential impact: New ways to access archives for archeology, history, or architecture, …

Motivation

What if we could automatically learn from this visual data?

Extract statistics of human behaviors across a city over time

"crossing street"

"bicycle accident"

"riding bicycle"

Potential impact: new ways to optimize road safety, urban planning or commerce in cities

Motivation

What if we could automatically learn from this visual data?

Learn to localize and navigate in changing conditions.

[Sattler et al., CVPR 2018]

[Taira et al., CVPR 2018]

How to recognize a chair?

Problem: Hard to design visual representation by hand






How to define the appearance of a chair?

Supervised machine learning

Positive examples (chairs)

Negative examples (other objects)

Training data

$f($  $) = +1$

$f($  $) = -1$

Image classifier

Mark I Perceptron [Rosenblatt'57]

Change parameters of f to minimize # of errors on training data.

Training procedure

Supervised machine learning

Positive examples $y_i = 1$

Negative examples $y_i = -1$

Training data

$\{x_i, y_i\}$

Images Labels (+-1)

$f($  $) = +1$

$f($  $) = -1$

Image classifier

Error (loss) on the all training data

Regularization

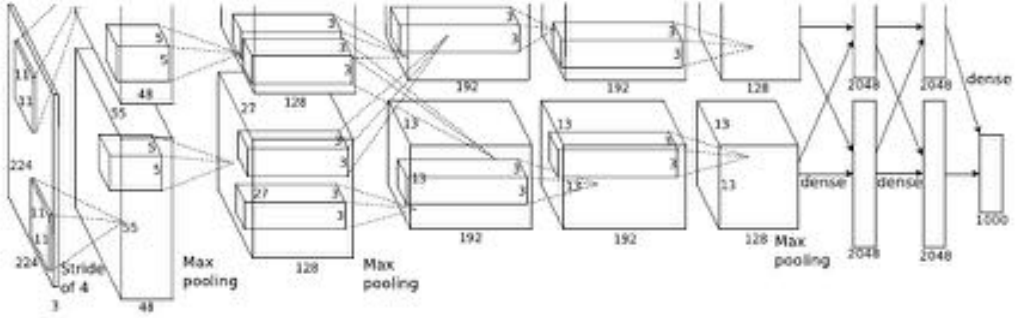
$$\min_f \sum_{i=1}^{N} \ell\left(y_i, f(x_i)\right) + \Omega(f)$$

Error on one example

Training procedure

Supervised machine learning: in practice

Millions of annotated
training examples
[from the Internet]



Classifier with
millions of parameters

Powerful training hardware
Days to weeks of training

Limitation I: Can we annotate the entire visual world?

Problem: annotation is expensive and can introduce biases

Currently: tedious **manual annotation**

Annotation is often ambiguous:
Table? / Dining Table? / Desk? / Bench?

Limitation II: What is the right granularity of visual representation?

Problem: the “visual vocabulary” is large, a priori unknown and task dependent

What is the set of **manipulation actions** that can be done with a particular **tool**?

What is the set of human behaviors that correlate with **pedestrian accidents**?

Solution: learn without human supervision
[Self-supervised learning]

Unsupervised learning
[Self-supervised learning]

Weakly-supervised learning

Learn from available meta-data :

e.g. video + *text, speech, audio, …*

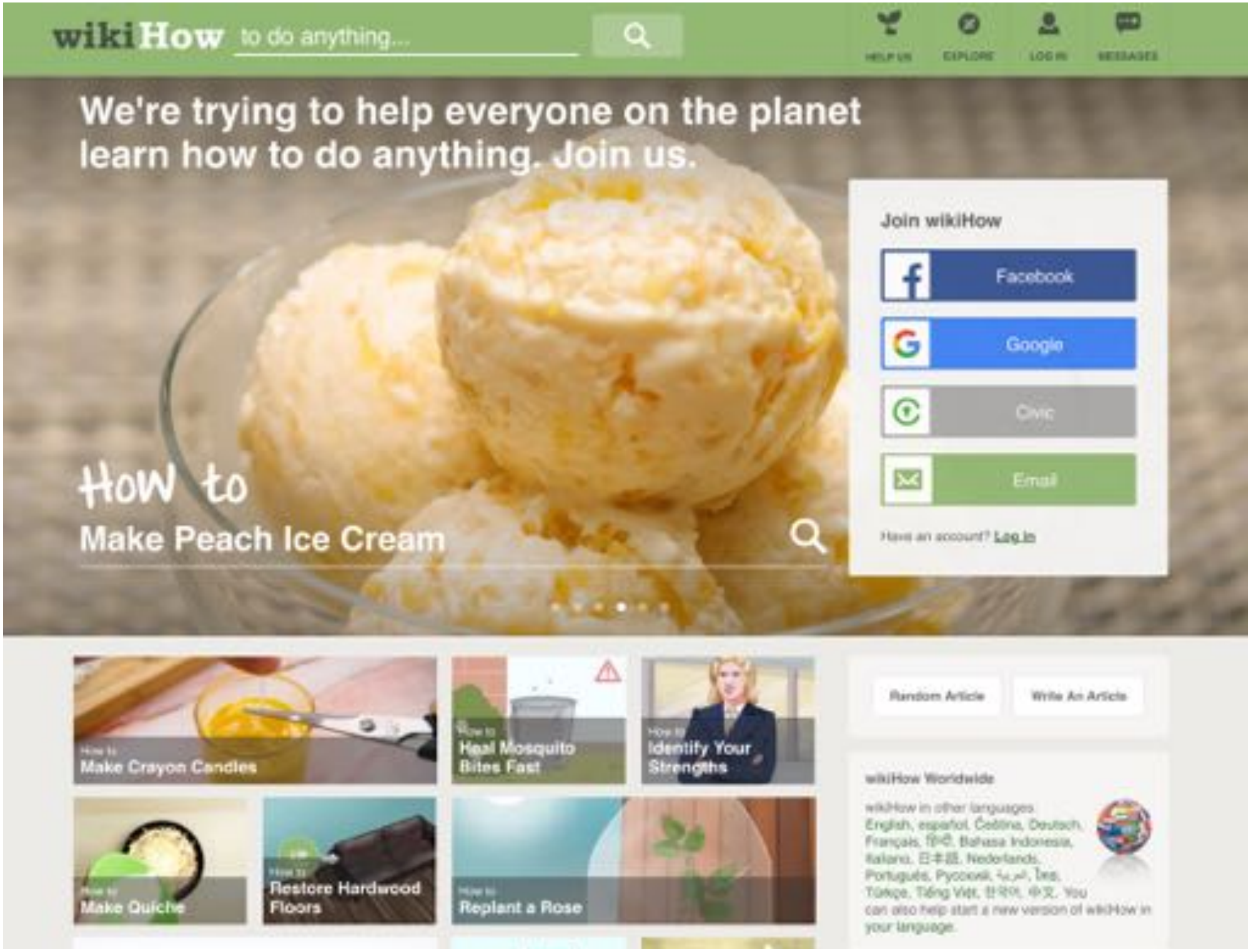
[Multi-modal self-supervised learning]

Learning by interaction with environment
[Reinforcement learning]

See also: https://gidariss.github.io/self-supervised-learning-cvpr2021

Example of meta-data: narrated instructional videos

[Alyarac et al., CVPR 2016]


wikiHow
to do anything...
HELP US
EXPLORE
LOG IN
MESSAGES
We're trying to help everyone on the planet learn how to do anything. Join us.
How to
Make Peach Ice Cream
Join wikiHow
Facebook
Google
Civic
Email
Have an account? Log in
How to
Make Crayon Candles
How to
Heal Mosquito Bites Fast
How to
Identify Your Strengths
Random Article
Write An Article
How to
Make Quiche
How to
Restore Hardwood Floors
How to
Replant a Rose
wikiHow Worldwide
wikiHow in other languages:
English, español, Čeština, Deutsch, Français, हिन्दी, Bahasa Indonesia, Italiano, 日本語, Nederlands, Português, Русский, العربية, ไทย, Türkçe, Tiếng Việt, 한국어, 中文. You can also help start a new version of wikiHow in your language.

Going WikiHow scale – the HowTo100M dataset

23K tasks • 1.3M videos • 130M clip-caption pairs

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]
[Miech, Alayrac, Smaira, Laptev, Sivic, Zisserman, CVPR 2020]

Going WikiHow scale

HowTo100M dataset

Dataset	Clips	Captions	Videos	Duration	Source	Year
Charades [42]	10k	16k	10,000	82h	Home	2016
MSR-VTT [52]	10k	200k	7,180	40h	Youtube	2016
YouCook2 [61]	14k	14k	2,000	176h	Youtube	2018
EPIC-KITCHENS [5]	40k	40k	432	55h	Home	2018
DiDeMo [11]	27k	41k	10,464	87h	Flickr	2017
M-VAD [46]	49k	56k	92	84h	Movies	2015
MPII-MD [37]	69k	68k	94	41h	Movies	2015
ANet Captions [22]	100k	100k	20,000	849h	Youtube	2017
TGIF [23]	102k	126k	102,068	103h	Tumblr	2016
LSMDC [38]	128k	128k	200	150h	Movies	2017
How2 [39]	185k	185k	13,168	298h	Youtube	2018
HowTo100M	**136M**	**136M**	**1.221M**	**134,472h**	Youtube	2019

23K tasks • 1.3M videos • 130M clip-caption pairs

Learn joint text-video embedding

Given a set of inputs x_i and supervisory meta-data y_i, $i = 1,\ldots,N$ learn embeddings $f(x_i)$ and $g(y_i)$ by solving

$$\min_{f,g} \sum_{i=1}^{N} \underbrace{\ell(g(y_i), f(x_i))}_{\text{Discriminative loss on data}} + \underbrace{\Omega(f,g)}_{\text{Regularization}}$$

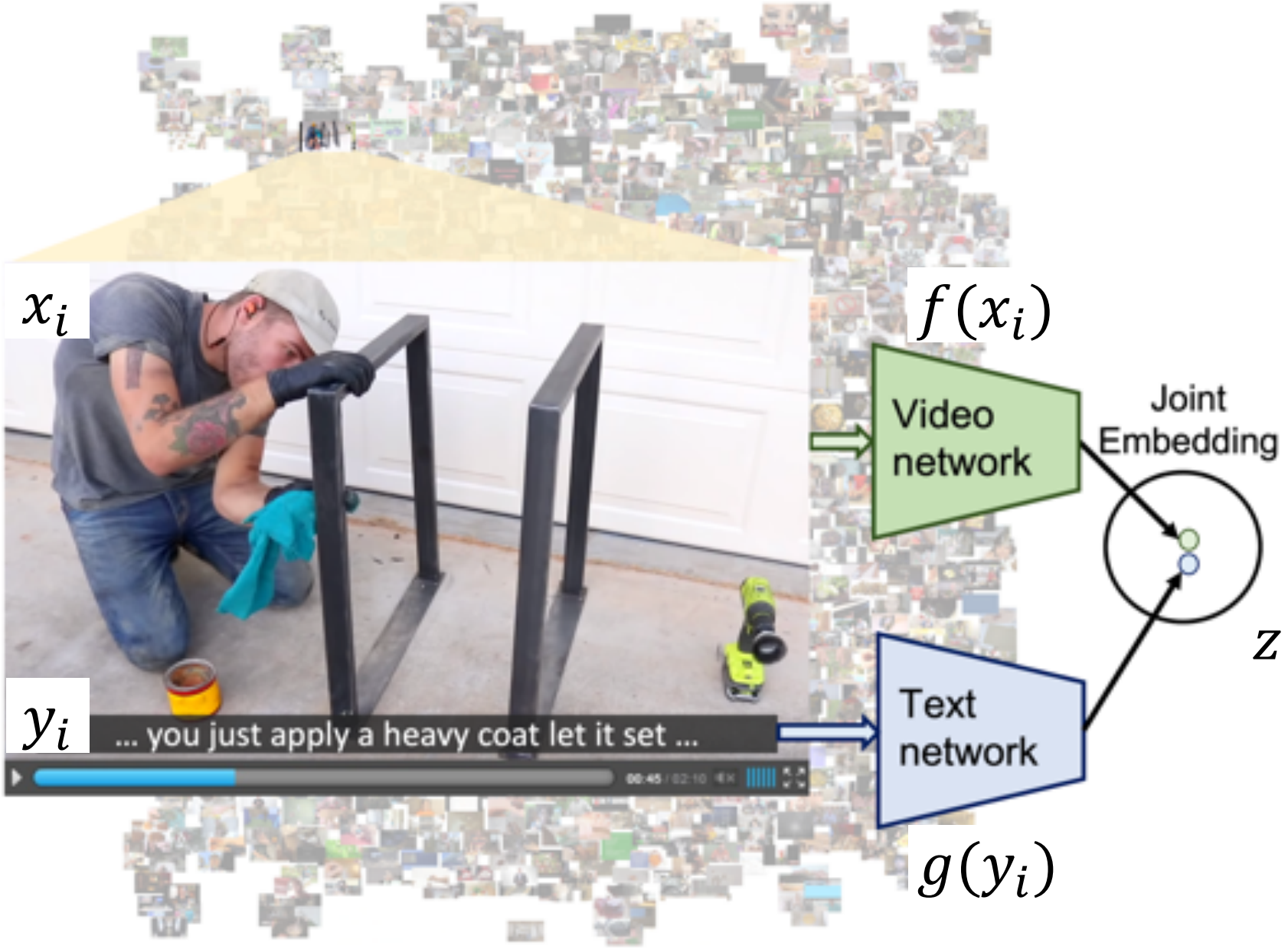
Scientific challenges:

- What is the appropriate form of these mappings and the loss?
- How to learn the mappings from the weak and noisy supervision?

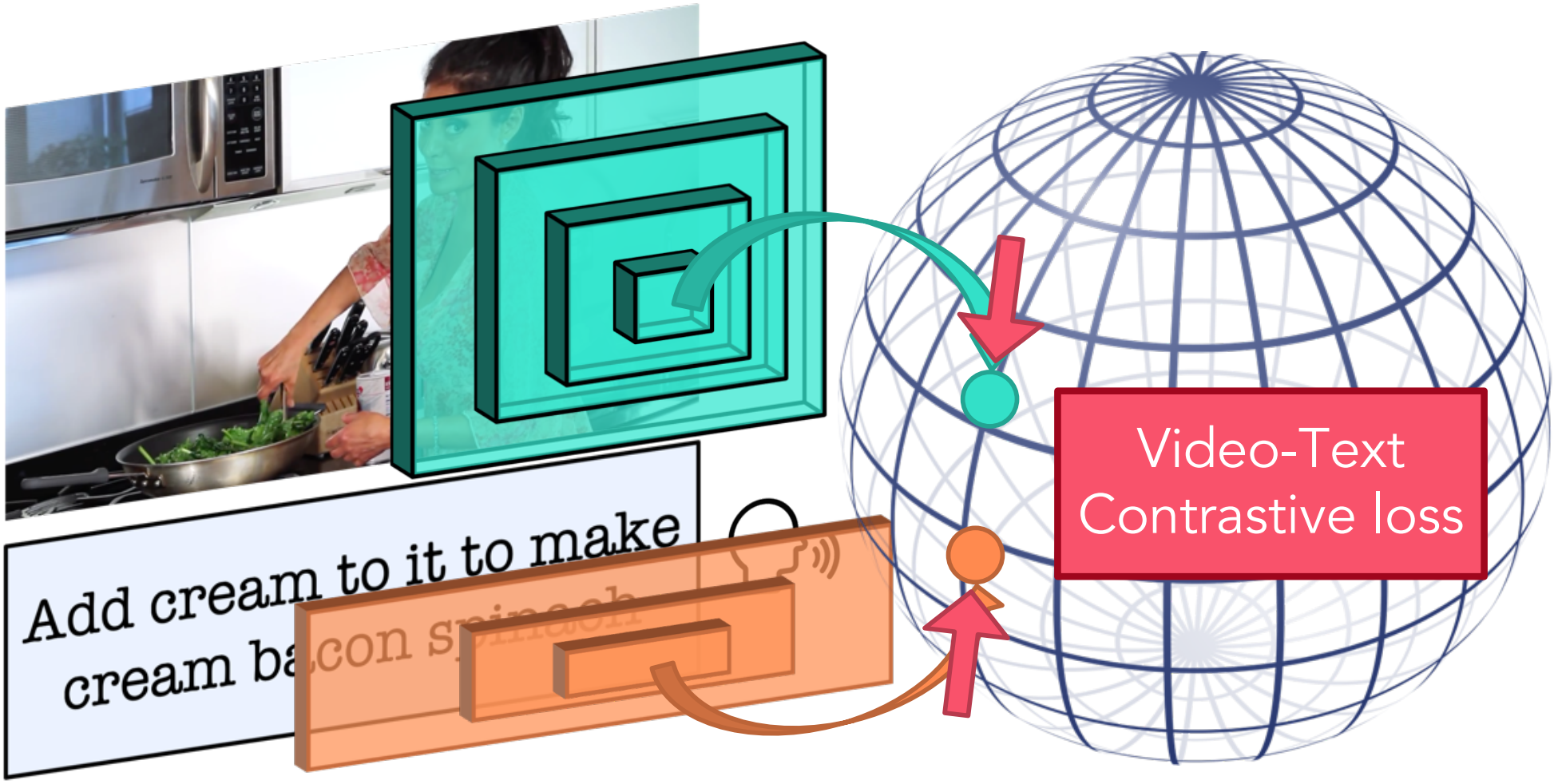
[Gong et al., 2013; Mikolov et al., 2013; Weston et al., 2011; Frome et al., 2013]

Also recently (large-scale learning from still images with text):
Open-AI CLIP: [Radford et al., Learning Transferable Visual Models From Natural Language Supervision, https://arxiv.org/abs/2103.00020]
Google ALIGN: [Jia et al., Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, https://arxiv.org/abs/2102.05918]

Learn joint text-video embedding



End-to-End Learning of Visual Representations from Uncurated Instructional Videos



[Miech, Alayrac, Smaira, Laptev, Sivic, Zisserman, CVPR 2020, https://arxiv.org/abs/1912.06430]

Speech Recognition

spinachs what's the name

keep it simple you just want to add

fresh herbs maybe some oregano

you can add cilantro basil they give

give it a couple more tosses

Time

spinachs what's the name

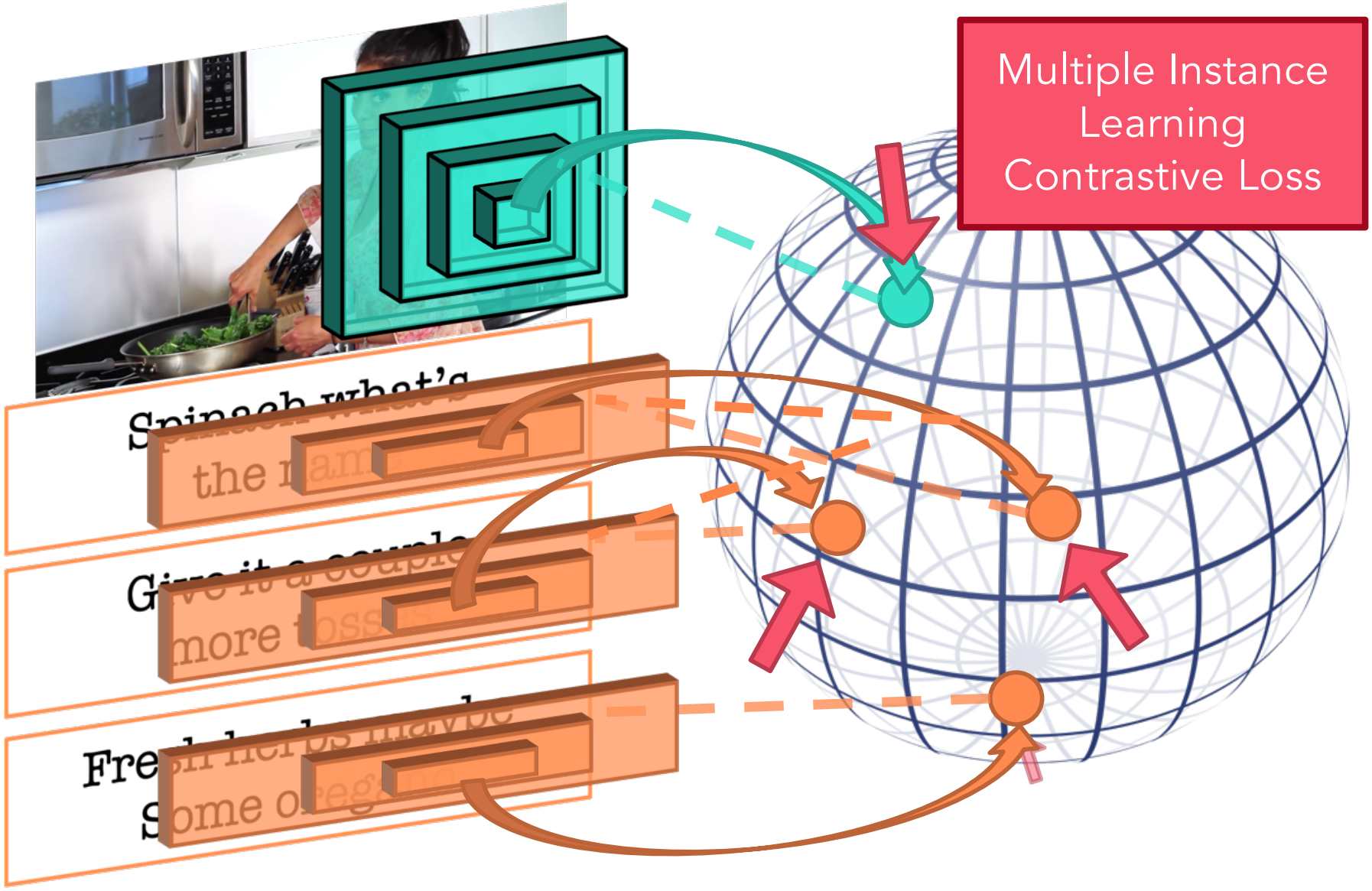
keep it simple you just want to add

fresh herbs maybe some oregano

you can add cilantro basil they give

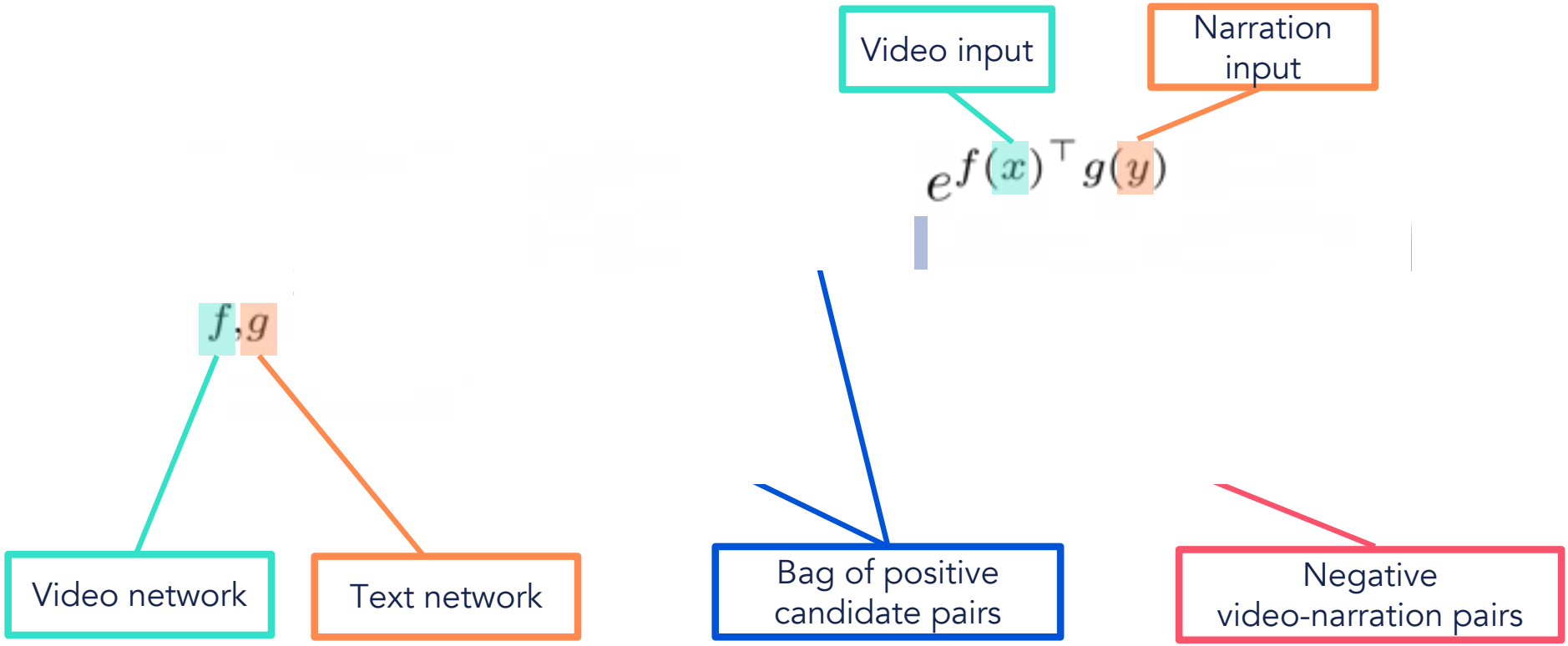
give it a couple more tosses

Multiple-Instance Learning Objective



Our formulation: MIL-NCE

Multiple Instance Learning - Noise Contrastive Estimation



Our formulation: MIL-NCE

Multiple Instance Learning - Noise Contrastive Estimation

Spinach what's the name

Give it a couple more tosses

Fresh herbs maybe Some oregano

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

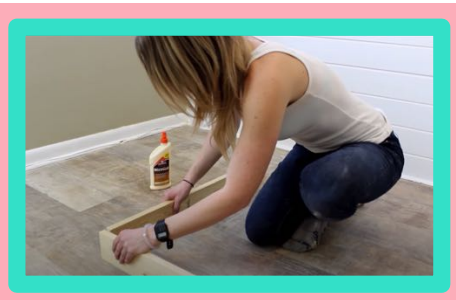
Bag of positive candidate pairs

Our formulation: MIL-NCE

Multiple Instance Learning - Noise Contrastive Estimation

Let's glue the piece of woods

Keep it simple you Just want to add

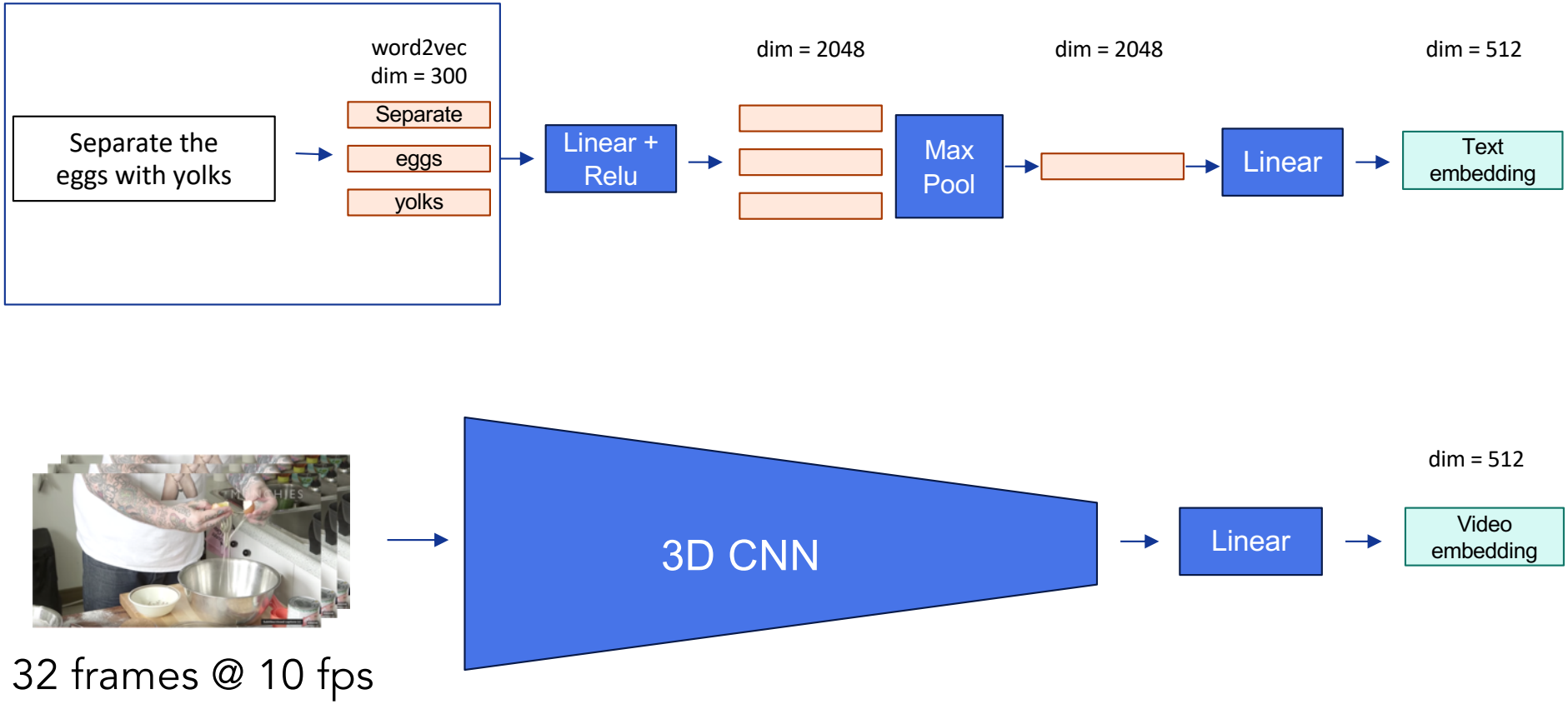


Fresh herbs maybe Some oregano

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

Negative video-narration pairs

Video-Text Model Architecture



[Miech, Alayrac, Smaira, Laptev, Sivic, Zisserman, CVPR 2020, https://arxiv.org/abs/1912.06430]

Evaluation: the downstream tasks

Action recognition

Action Localization

Action Video retrieval

HMDB-51

MSR-VTT

uTub

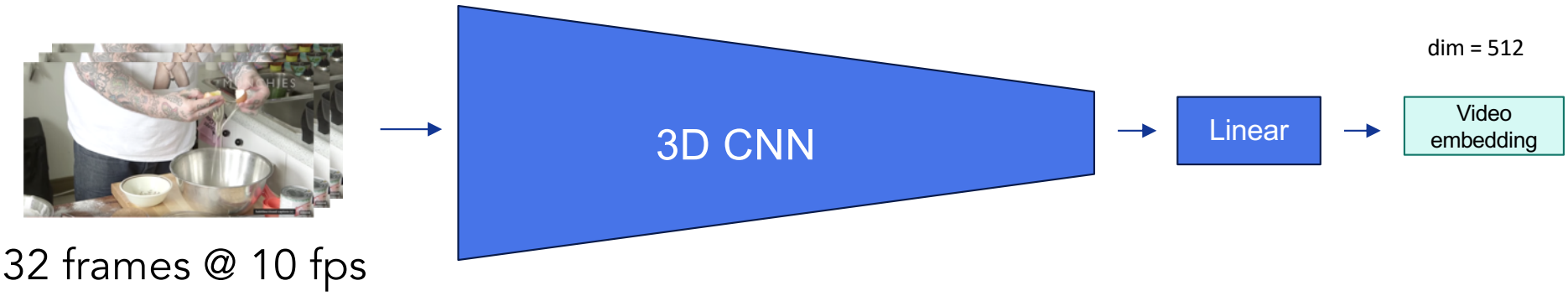
Segments

UCF-101

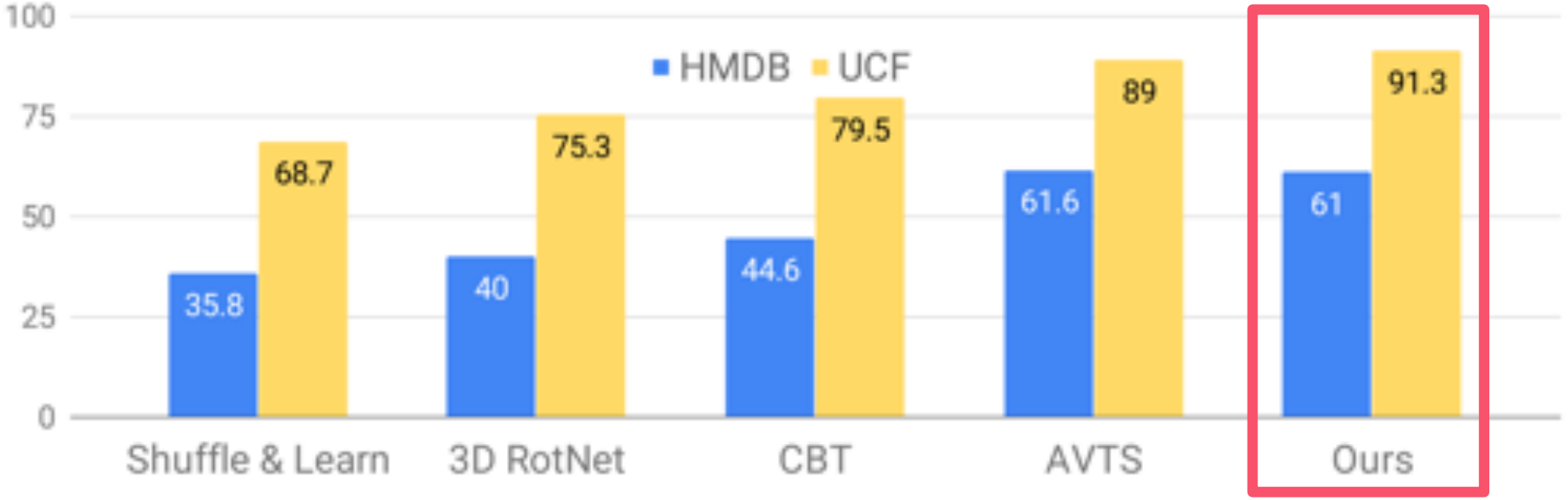
YouCook2

Evaluation: experimental set-ups

- Linear probe / Nearest neighbour with fixed representation
- Finetuning on target task and dataset
- Text to video retrieval



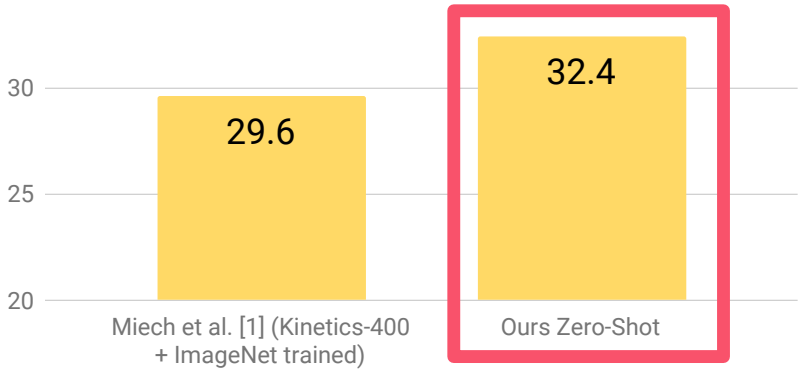
Action recognition:
On UCF-101 and HMDB-51, state-of-the-art results compared **to self-supervised video representations.**

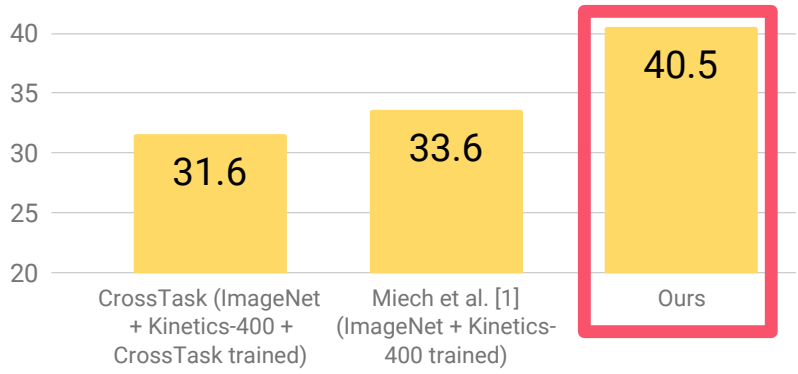


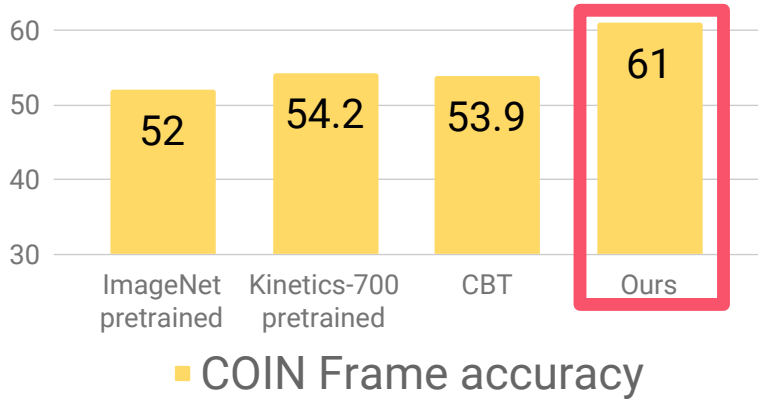
[Miech, Alayrac, Smaira, Laptev, Sivic, Zisserman, CVPR 2020, https://arxiv.org/abs/1912.06430]

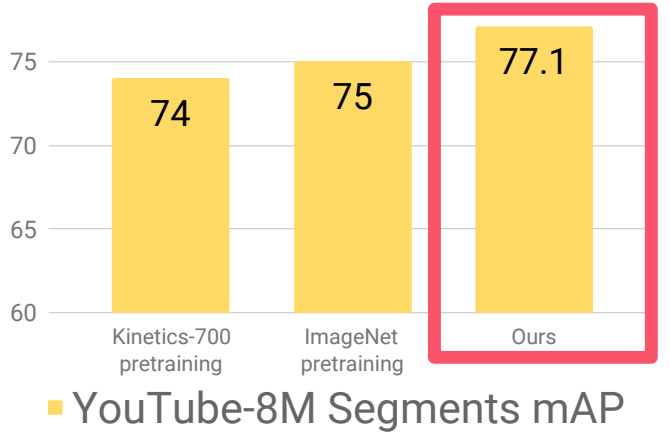
Results can be improved using also the audio signal. See the later part of the lecture.

The representation **outperforms fully-supervised ones** on several downstream tasks.







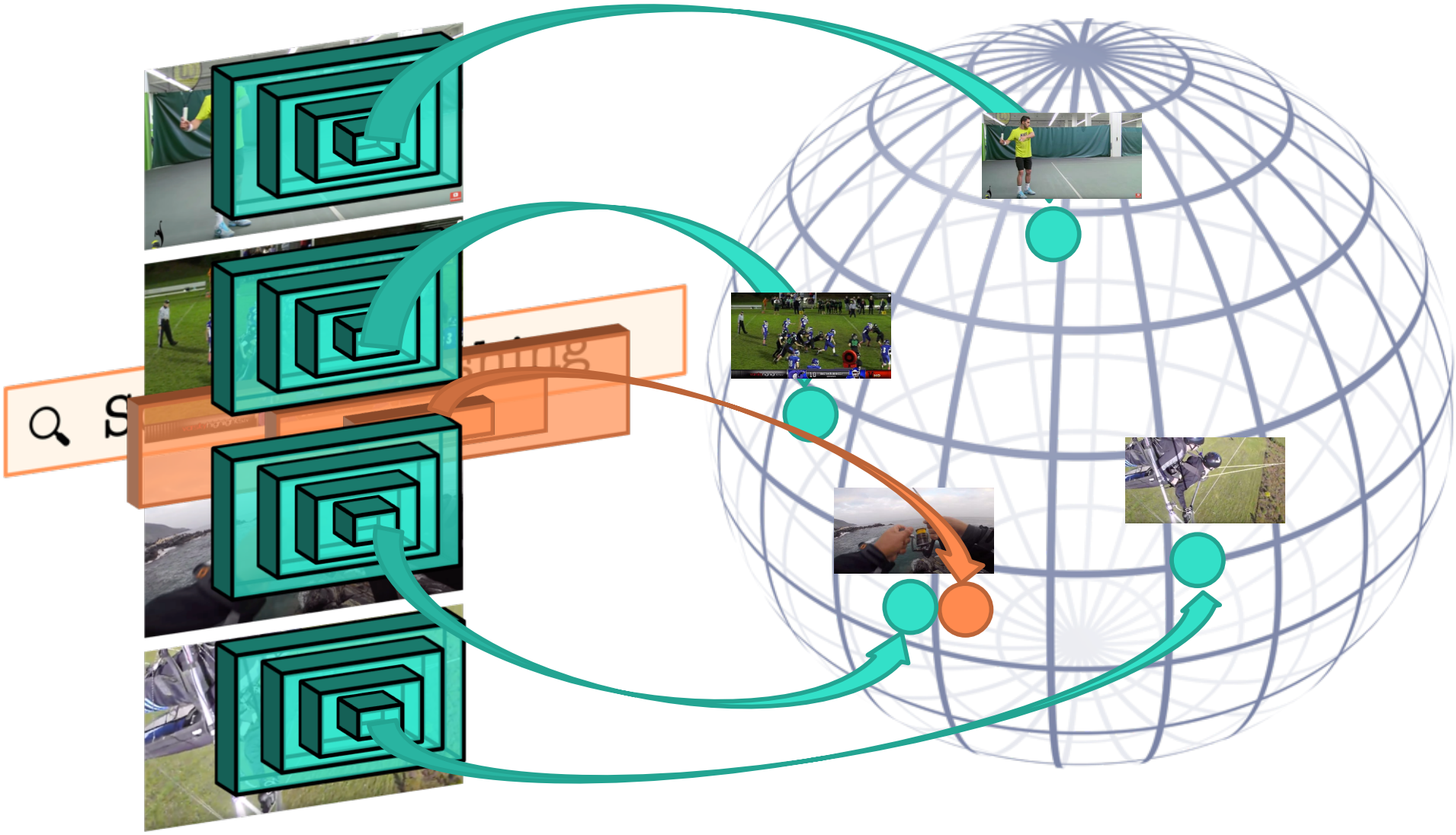


[1] *A.* Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic,
HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, in ICCV, 2019.

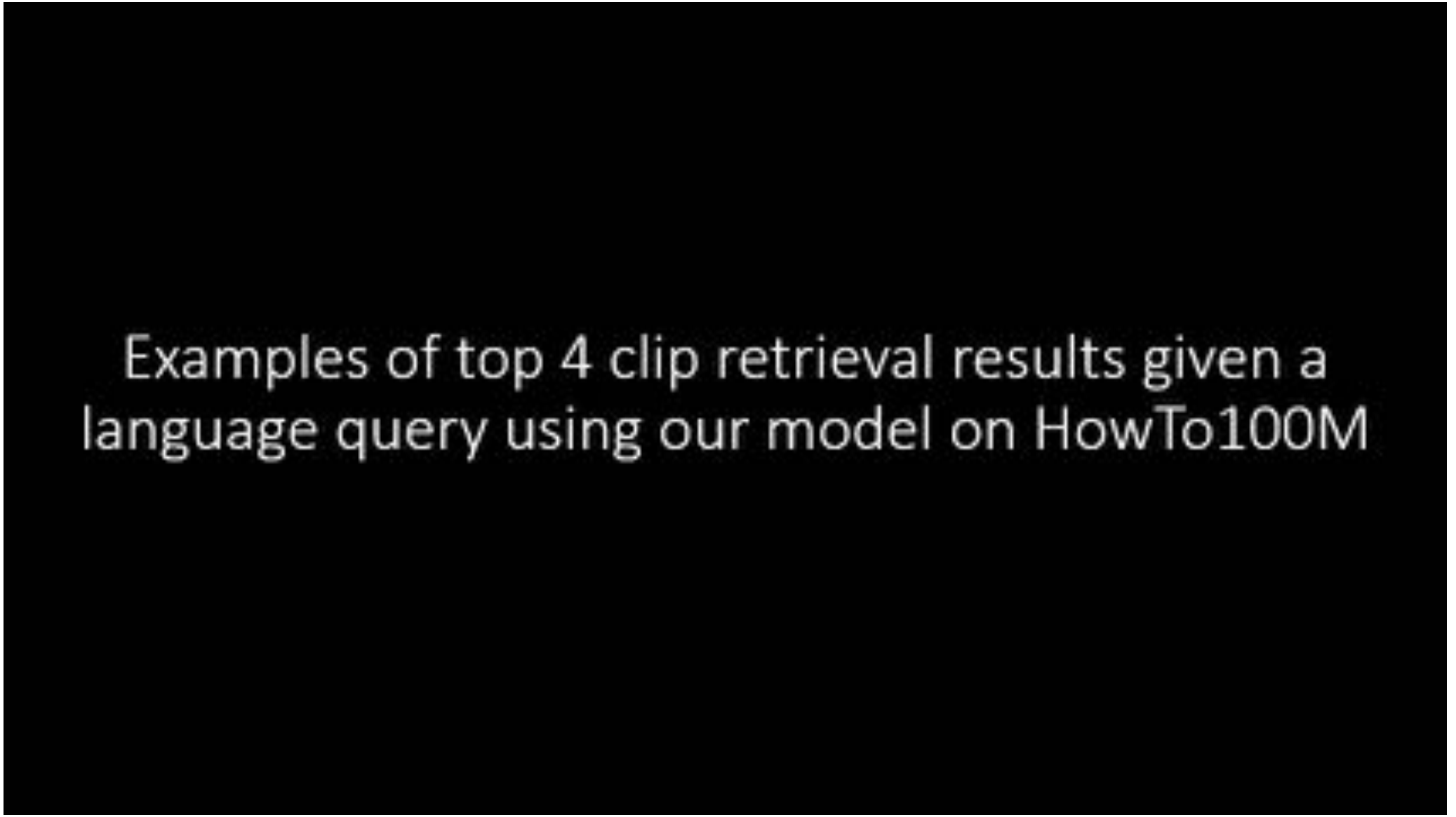
[Miech, Alayrac, Smaira, Laptev, Sivic, Zisserman, CVPR 2020, https://arxiv.org/abs/1912.06430]

Evaluating the joint text-video representation

Text-to-Video retrieval evaluation



Text to video retrieval



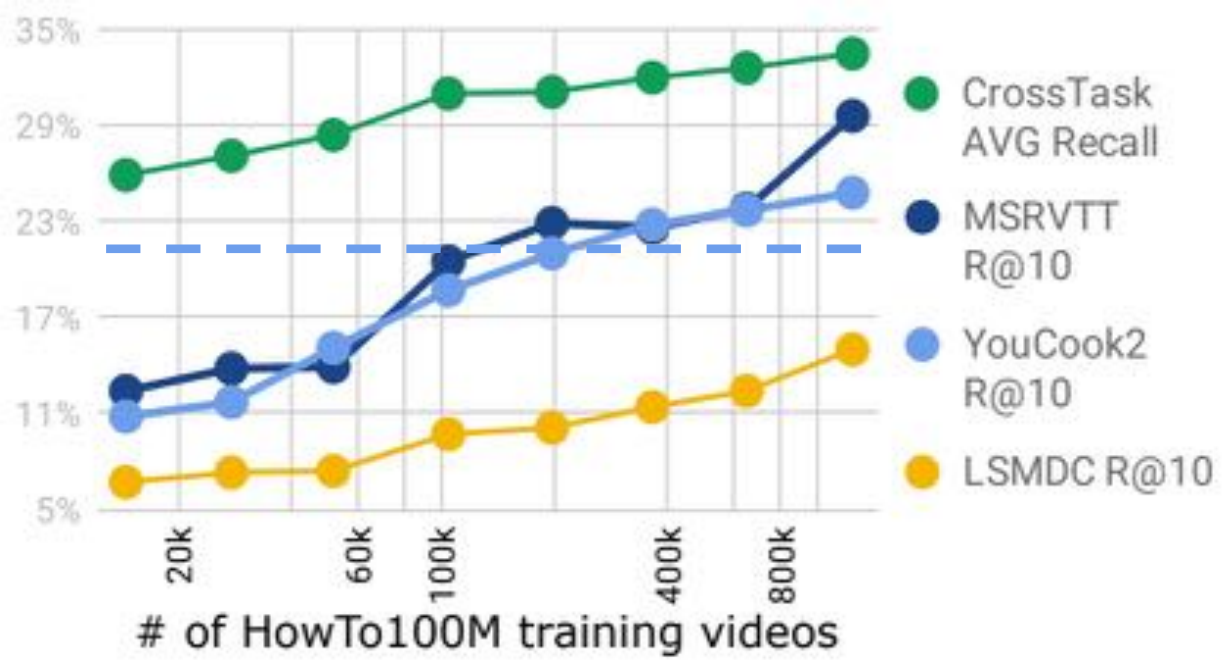
[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019, https://arxiv.org/abs/1906.03327]

Results: Text-to-video retrieval

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019, https://arxiv.org/abs/1906.03327]

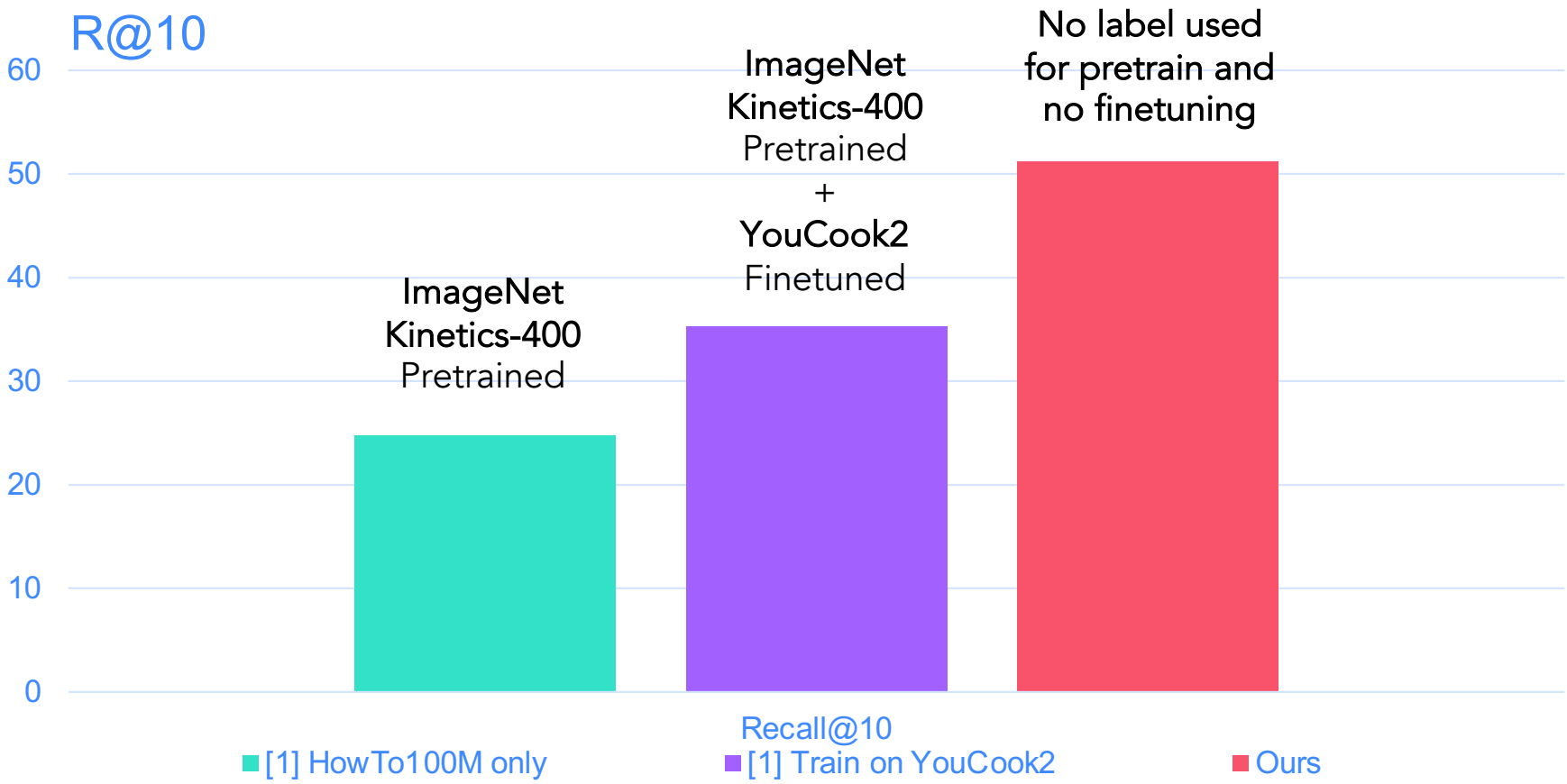
Results: Text-to-video retrieval

Fully supervised with manual annotations



[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019, https://arxiv.org/abs/1906.03327]

YouCook2 Zero-Shot Text-to-Video retrieval



[1] *A.* Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic,
HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, in ICCV, 2019.

[Miech, Alayrac, Smaira, Laptev, Sivic, Zisserman, CVPR 2020, https://arxiv.org/abs/1912.06430]

www.di.ens.fr/willow/research/mil-nce

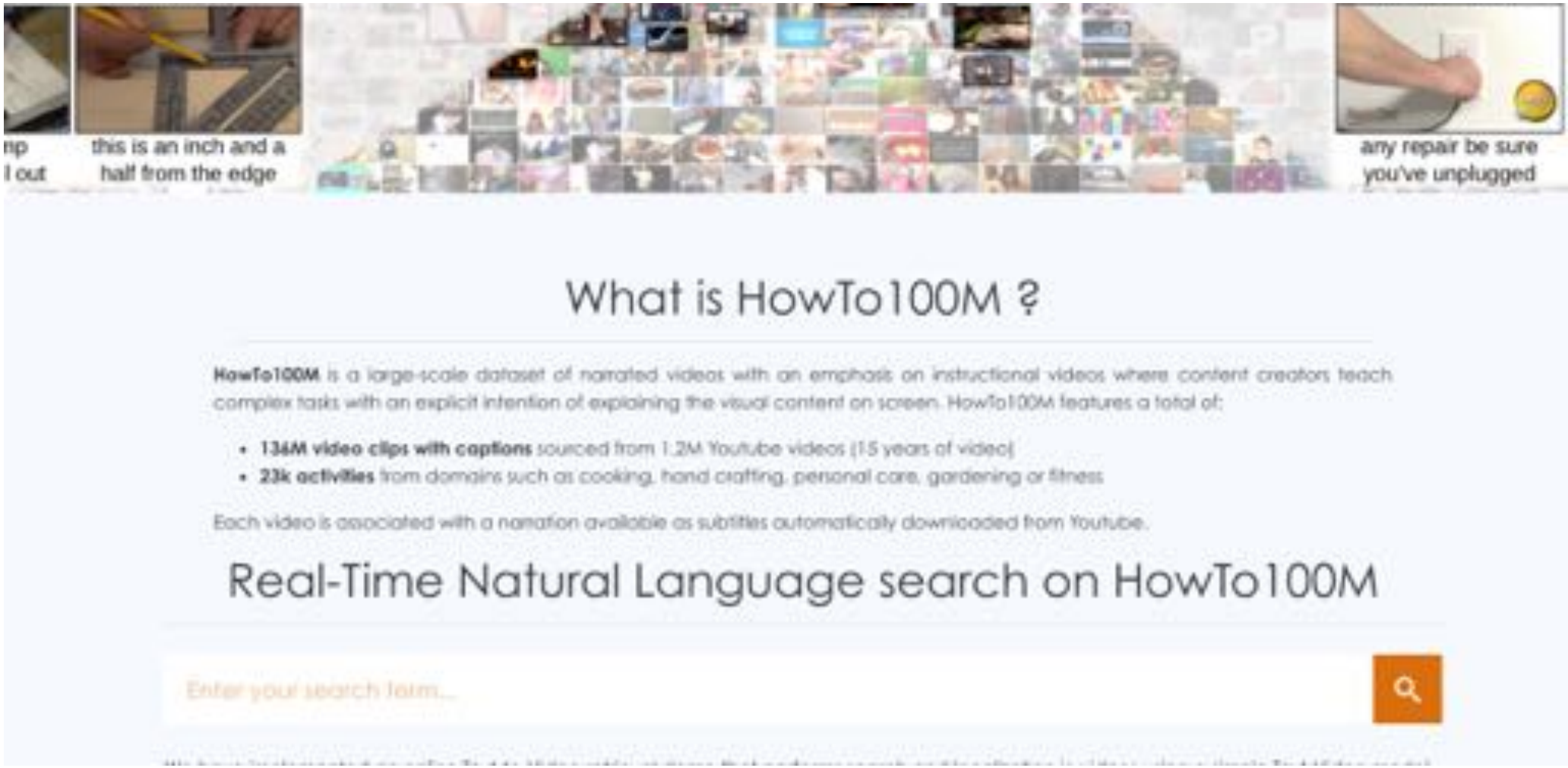
Abstract Search demo Download Team

End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Annotating videos is cumbersome, expensive and not scalable. Yet, many strong video models still rely on manually annotated data. With the recent introduction of the HowTo100M dataset, narrated videos now offer the possibility of learning video representations without manual supervision. In this work we propose a new learning approach, MIL-NCE, capable of addressing misalignments inherent to narrated videos. With this approach we are able to learn strong video representations from scratch,

Code, models, data and **demo** available online

https://www.di.ens.fr/willow/research/howto100m/
https://www.di.ens.fr/willow/research/mil-nce/



[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019, https://arxiv.org/abs/1906.03327]
[Miech, Alayrac, Smaira, Laptev, Sivic, Zisserman, CVPR 2020, https://arxiv.org/abs/1912.06430]

Another example of downstream task: Video question answering

Open-Ended Questions:
Where are the men?

Answer: **Track**

Multiple-Choice Questions:
What are the lined up men doing?

Proposal 1: Running

Proposal 2: Talking

Proposal 3: Shaving

[Yang, Miech, Sivic, Laptev, Schmid, 2021, https://arxiv.org/pdf/2012.00451.pdf]

Video question answering : challenges

- **Variability:** VideoQA requires the ability to recognize actions, objects, colors at different spatio-temporal granularities

- **Annotations:** Obtaining manually annotated VideoQA data is expensive and not scalable

Question: How many times does the cat lick?

Answer: **7 times**

Question: What does the cat do 3 times?

Answer: **puts head down**

Question: What is the color of the bulldog?

Answer: **brown**

Automatically generated HowToVQA69 dataset

Dataset	Number of videos	Number of questions
MovieQA [76]	6.8K	15K
PororoQA [39]	16K	9K
VideoQA [95]	18K	175K
YouTube2Text-QA [92]	2K	99K
MSRVTT-QA [86]	10K	244K
MSVD-QA [86]	2K	51K
TGIF-QA [34]	72K	165K
TVQA [44]	22K	152K
SVQA [70]	12K	118K
Social-IQ [94]	1.3K	8K
ActivityNet-QA [93]	5.8K	58K
KnowIT VQA[27]	12K	24K
How2QA [47]	9K	44K
DramaQA [16]	24K	18K
HowToVQA69M	**1.2M**	**69M**

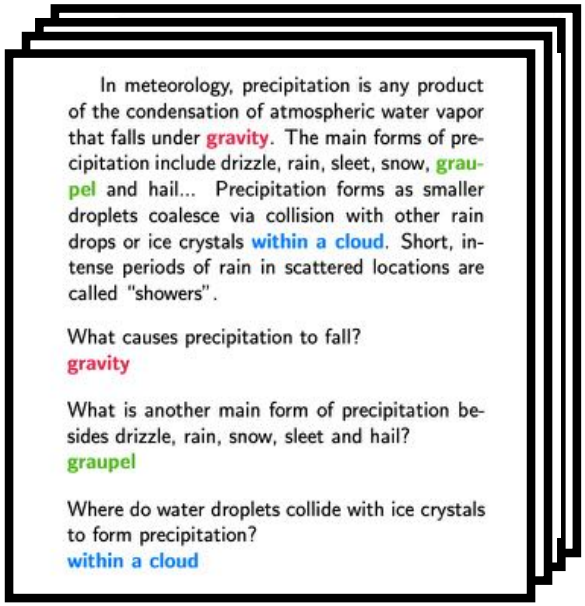
Weak supervision from narrated instructional videos

[Miech 2019] HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, A. Miech et al

Text-only supervision for automatic generation of VideoQA data

To generate VideoQA data, we rely on language models [Raffel 2020] trained on text-only annotations

Manually annotated QA text corpus



Training

Training

Answer extractor Transformer

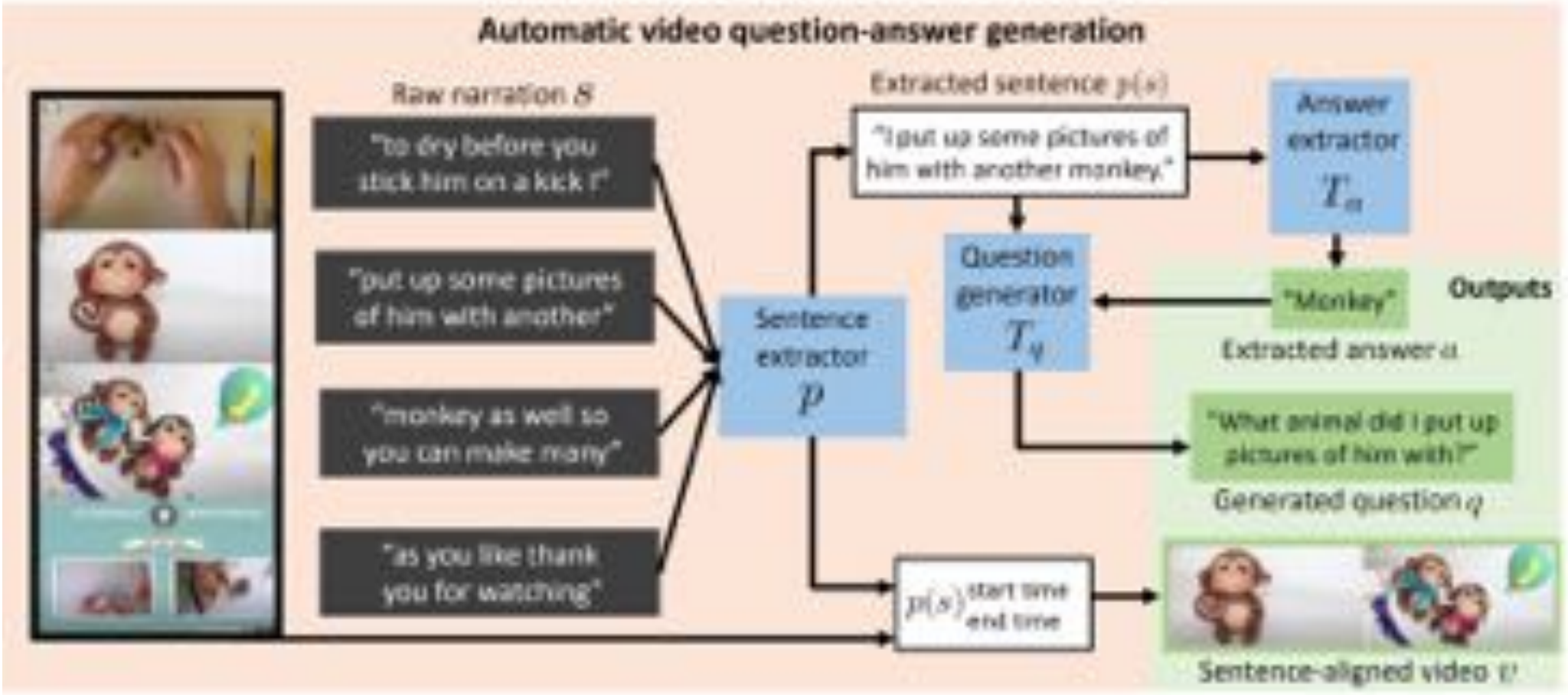
T_a

Question generator Transformer

T_q

[Raffel 2020] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, C. Raffel et al

Automatic large-scale generation of VideoQA data



Automatically generated HowToVQA69 dataset

ASR: And the last thing that goes on top would be the spinach.
Question: What is the last thing that goes on top?
Answer: Spinach

ASR: So I've got nine blobs of dough here a little bit sticky.
Question: How many blobs of dough are there?
Answer: Nine

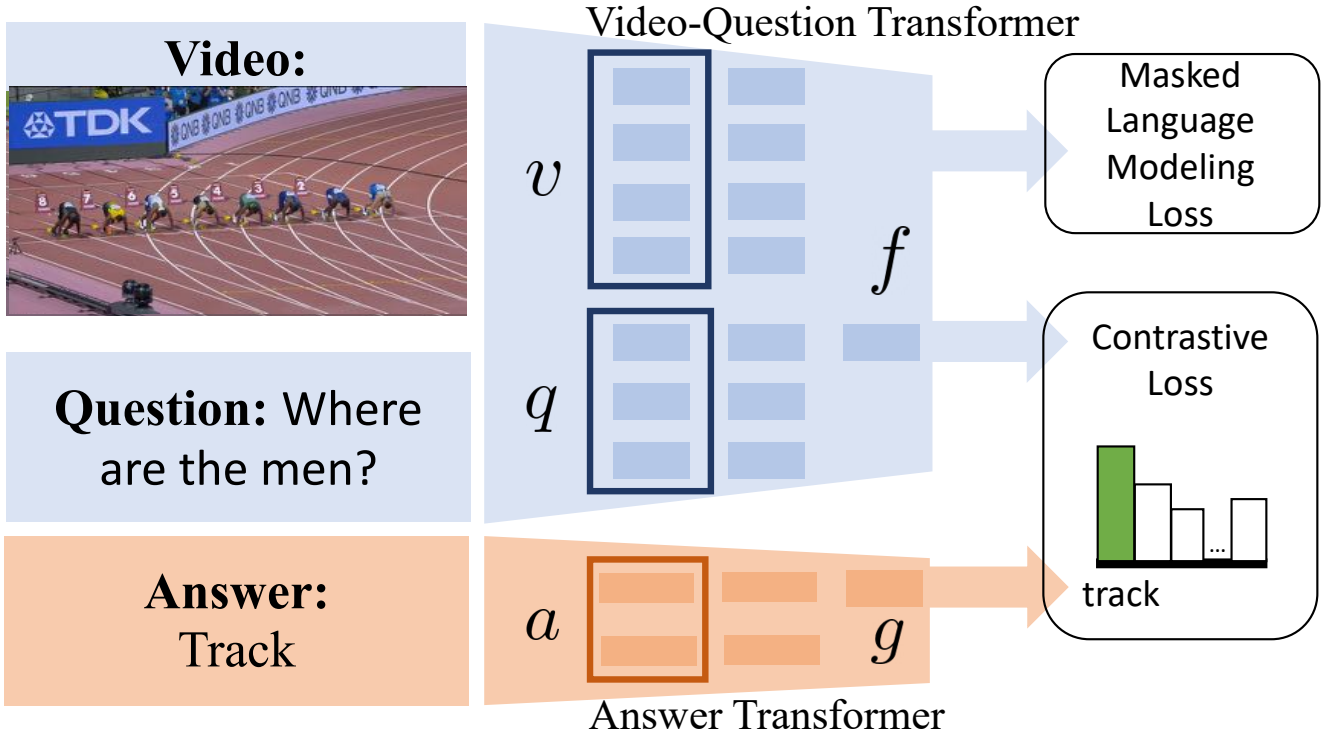
ASR: So you bring it to a point and we'll, just cut it off at the bottom.
Question: What do we do at the bottom?
Answer: Cut it off

ASR: Just let them do whatever they want and it'll still look pretty cool.
Question: What's the best way to make it look cool?
Answer: Let them do whatever they wants

ASR: The onions are chopped pretty much the same size.
Question: What are chopped pretty much the same size as the other vegetables?
Answer: The onions

ASR: ...I've had over a hundred emails.
Question: How many emails have I had?
Answer: over a hundred

VideoQA model (*VQA-T*) and training procedure on HowToVQA69M

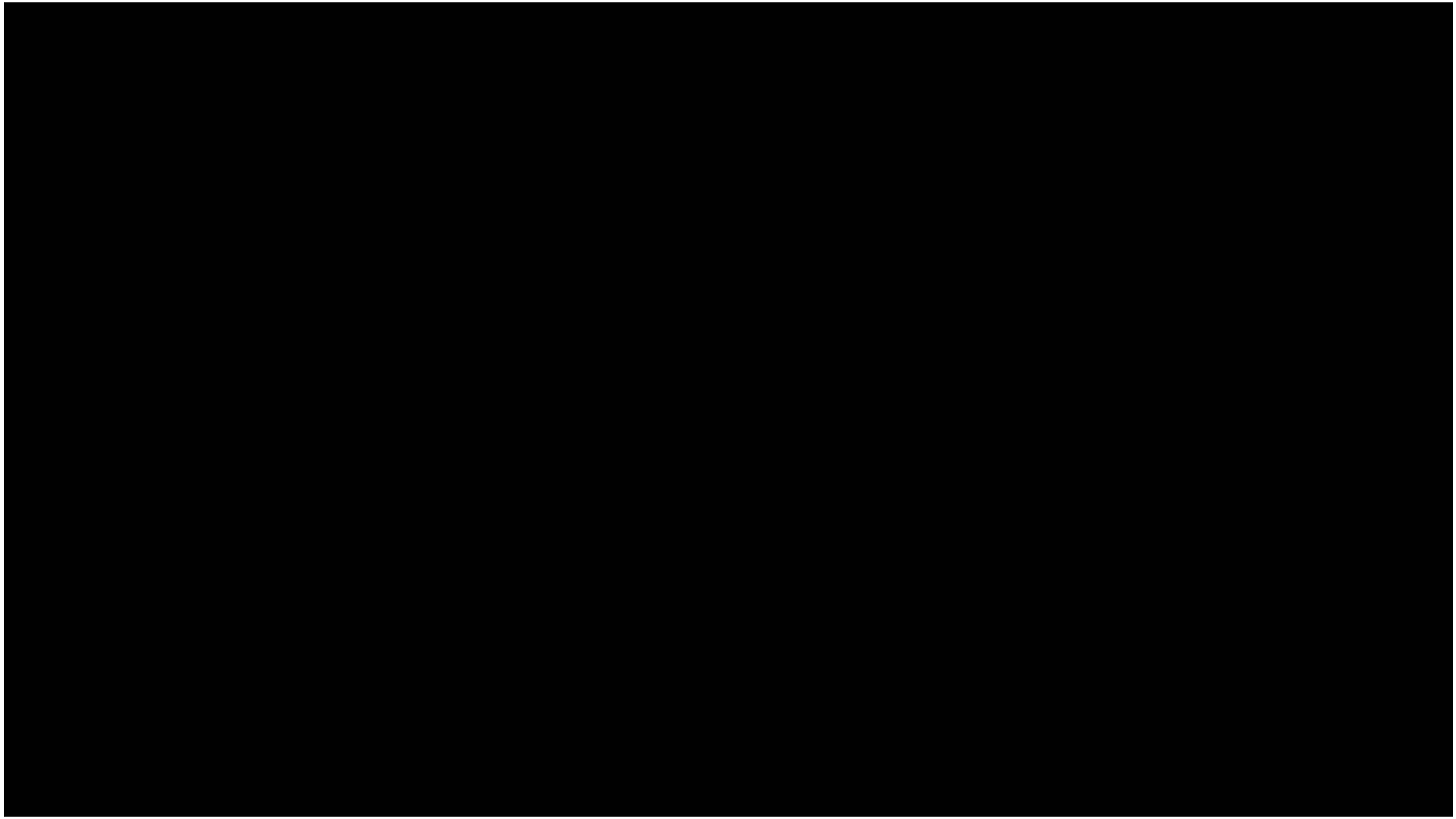


Results

Method	Pretraining data	MSRVTT-QA	MSVD-QA
E-SA [86]		29.3	27.6
ST-TP [34]		30.9	31.3
AMU [86]		32.5	32.0
Co-mem [26]		32.0	31.7
HME [22]		33.0	33.7
LAGCN [32]		—	34.3
HGA [36]		35.5	34.7
QueST [35]		34.6	36.1
HCRN [41]		35.6	36.1
Clip-BERT [43]	COCO [14]+ Visual Genome [40]	37.4	—
SSML [5]	HowTo100M	35.1	35.1
CoMVT [67]	HowTo100M	39.5	42.6
VQA-T	∅	39.6	41.2
VQA-T	HowToVQA69M	**41.5**	**46.3**

Table 5: Comparison with state of the art on MSRVTT-QA and MSVD-QA (top-1 accuracy).

Example results



Beyond language:

What about sound and audio?

Cross-modal contrastive learning

Self-Supervised MultiModal Versatile Networks

Jean-Baptiste Alayrac[1*] Adrià Recasens[1*] Rosalia Schneider[1*] Relja Arandjelović[1*]

Jason Ramapuram[2,3†] Jeffrey De Fauw[1] Lucas Smaira[1] Sander Dieleman[1]

Andrew Zisserman[1,4]

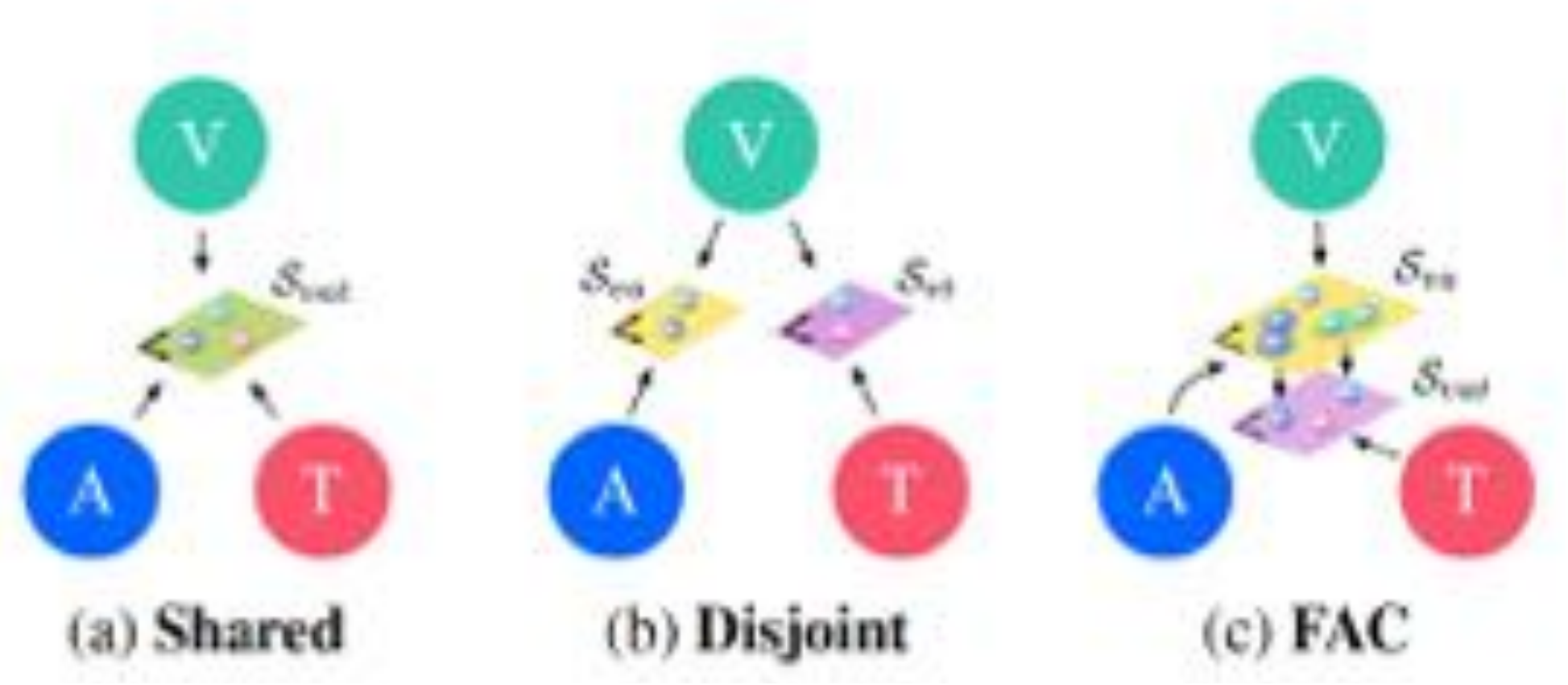
[1]DeepMind [2]Faculty of Science, Computer Science Dept., University of Geneva, HES-SO
[3]Geneva School of Business Administration (DMML Group)
[4]VGG, Dept. of Engineering Science, University of Oxford
{jalayrac, arecasens, rgschneider, relja}@google.com

[Alayrac et al., Self-Supervised MultiModal Versatile Networks, NeurIPS 2020]

Cross-modal contrastive learning

Learning: Video, Audio and Transcribed text from HowTo100M and Audio+Video from AudioSet

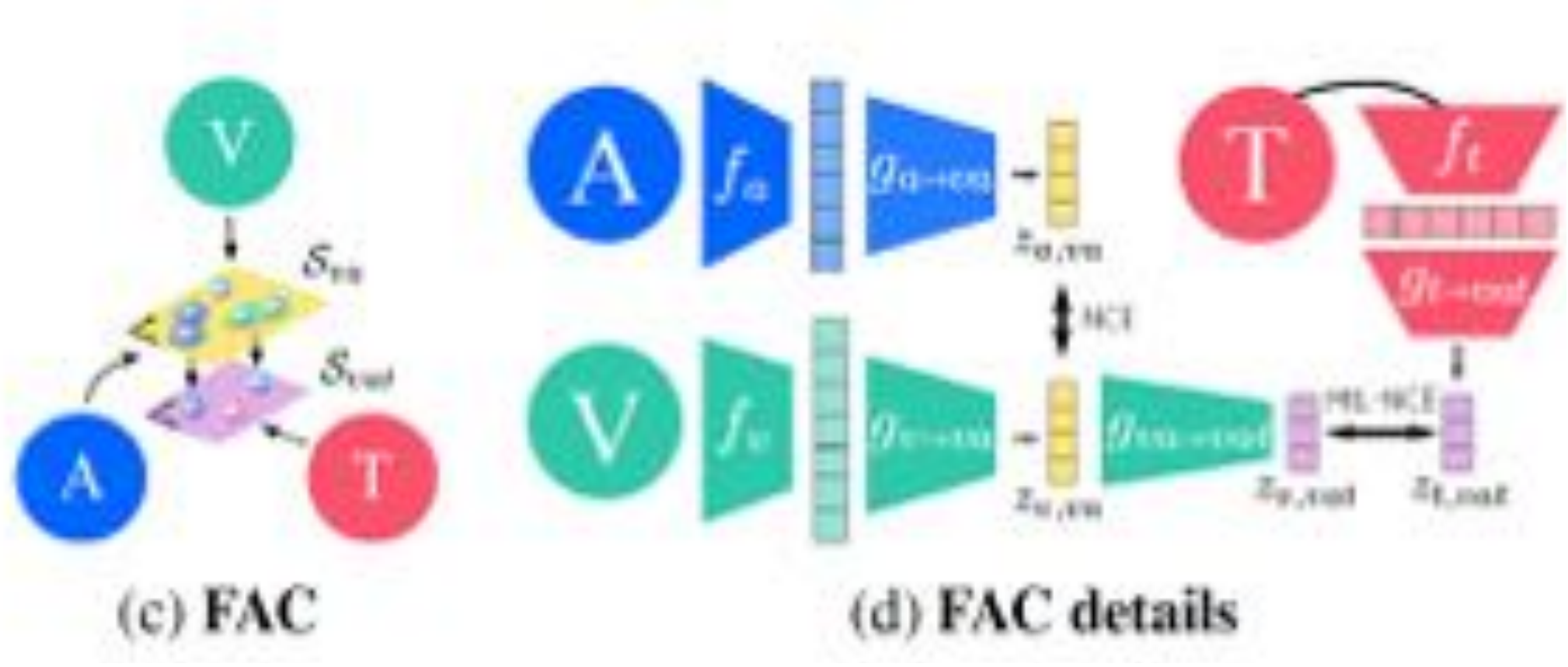
Explore different ways to combine modalities.



Learning from video, language and audio

Learning: Two **contrastive losses**: (1) audio and video (NCE), and (2) video and text (MIL-NCE).

Result: Three modalities better than two.



Results – Action Classification

Method	f_v (#params)	Train data	years	Mod.	UCF101 Linear	UCF101 FT	HMDB51 Linear	HMDB51 FT	ESC-50 Linear	AS MLP	K600 Linear
MIL-NCE [51]	I3D (12.1M)	HT	15	VT	83.4	89.1	54.8	59.2	/	/	
MIL-NCE [51]	S3D-G (9.1M)	HT	15	VT	82.7	91.3	53.1	61.0	/	/	
AVTS [43]	MC3 (11.7M)	AS	1	VA		89.0		61.6	80.6		
AVTS [43]	MC3 (11.7M)	SNet	1	VA					82.3		
AA+AV CC [34]	RN-50 (23.5M)	AS	1	VA						28.5	
CVRL [70]	R3D50 (33.3M)	K600	0.1	V							64.1
XDC [6]	R(2+1)D-18 (33.3M)	AS	1	VA		91.2		61.0	84.8		
XDC [6]	R(2+1)D-18 (33.3M)	IG65M	21	VA		94.2		67.4			
ELo [67]	R(2+1)D-50 (46.9M)	YT8M	13	VFA		93.8	64.5	67.4			
AVID [57]	R(2+1)D-50 (46.9M)	AS	1	VA		91.5		64.7	**89.2**		
GDT [64]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1	88.5		
GDT [64]	R(2+1)D-18 (33.3M)	IG65M	21	VA		**95.2**		72.8			
VA only (ours)	R(2+1)D-18 (33.3M)	AS	1	VA	83.9	91.5	60.0	70.1	85.6	29.7	55.5
VA only (ours)	S3D-G (9.1M)	AS	1	VA	84.7	90.1	60.4	68.2	86.1	29.7	59.8
VA only (ours)	S3D-G (9.1M)	AS+HT	16	VA	86.2	91.1	61.5	68.3	87.2	30.6	59.8
MMV FAC (ours)	S3D-G (9.1M)	AS+HT	16	VAT	89.6	92.5	62.6	69.6	87.7	30.3	68.0
MMV FAC (ours)	TSM-50 (23.5M)	AS+HT	16	VAT	91.5	94.9	66.7	73.2	86.4	30.6	67.8
MMV FAC (ours)	TSM-50x2 (93.9M)	AS+HT	16	VAT	**91.8**	**95.2**	**67.1**	**75.0**	88.9	**30.9**	**70.5**
Supervised [21, 42, 67, 74, 90]						96.8	71.5	75.9	86.5†	43.9	81.8

Results – Action Classification

Method	f_v (#params)	Train data	years	Mod.	UCF101 Linear	UCF101 FT	HMDB51 Linear	HMDB51 FT	ESC-50 Linear	AS MLP	K600 Linear
MIL-NCE [51]	I3D (12.1M)	HT	15	VT	83.4	89.1	54.8	59.2	/	/	
MIL-NCE [51]	S3D-G (9.1M)	HT	15	VT	82.7	91.3	53.1	61.0	/	/	
AVTS [43]	MC3 (11.7M)	AS	1	VA		89.0		61.6	80.6		
AVTS [43]	MC3 (11.7M)	SNet	1	VA					82.3		
AA+AV CC [34]	RN-50 (23.5M)	AS	1	VA						28.5	
CVRL [70]	R3D50 (33.3M)	K600	0.1	V							64.1
XDC [6]	R(2+1)D-18 (33.3M)	AS	1	VA		91.2		61.0	84.8		
XDC [6]	R(2+1)D-18 (33.3M)	IG65M	21	VA		94.2		67.4			
ELo [67]	R(2+1)D-50 (46.9M)	YT8M	13	VFA		93.8	64.5	67.4			
AVID [57]	R(2+1)D-50 (46.9M)	AS	1	VA		91.5		64.7	**89.2**		
GDT [64]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1	88.5		
GDT [64]	R(2+1)D-18 (33.3M)	IG65M	21	VA		**95.2**		72.8			
VA only (ours)	R(2+1)D-18 (33.3M)	AS	1	VA	83.9	91.5	60.0	70.1	85.6	29.7	55.5
VA only (ours)	S3D-G (9.1M)	AS	1	VA	84.7	90.1	60.4	68.2	86.1	29.7	59.8
VA only (ours)	S3D-G (9.1M)	AS+HT	16	VA	86.2	91.1	61.5	68.3	87.2	30.6	59.8
MMV FAC (ours)	S3D-G (9.1M)	AS+HT	16	VAT	89.6	92.5	62.6	69.6	87.7	30.3	68.0
MMV FAC (ours)	TSM-50 (23.5M)	AS+HT	16	VAT	91.5	94.9	66.7	73.2	86.4	30.6	67.8
MMV FAC (ours)	TSM-50x2 (93.9M)	AS+HT	16	VAT	**91.8**	**95.2**	**67.1**	**75.0**	88.9	**30.9**	**70.5**
Supervised [21, 42, 67, 74, 90]						96.8	71.5	75.9	86.5†	43.9	81.8

Cross-modal clustering: Audio-video only

Self-Supervised Learning by Cross-Modal Audio-Video Clustering

Humam Alwassel[1*] humam.alwassel@kaust.edu.sa

Dhruv Mahajan[2] dhruvm@fb.com

Bruno Korbar[2] bkorbar@fb.com

Lorenzo Torresani[2] torresani@fb.com

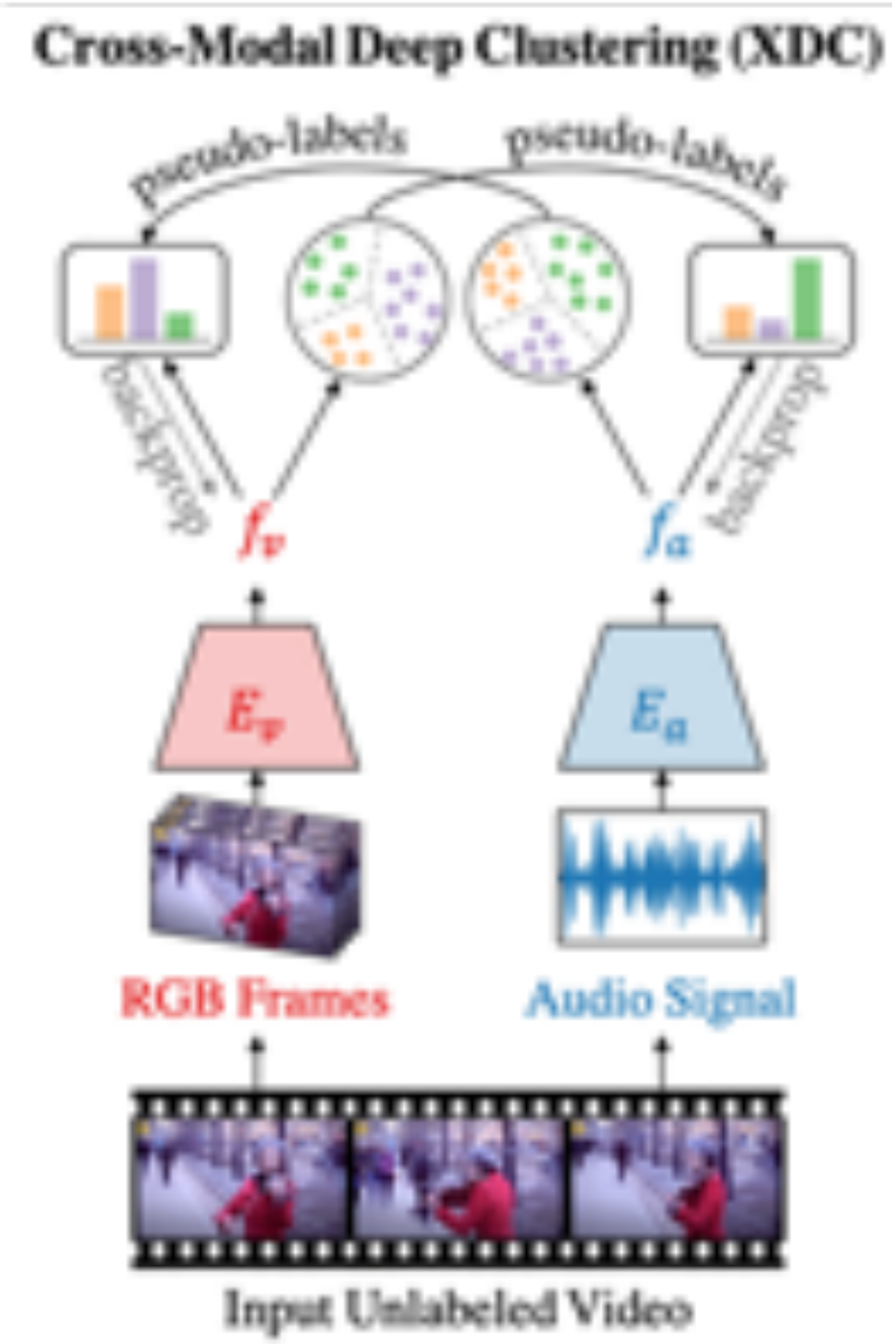
Bernard Ghanem[1] bernard.ghanem@kaust.edu.sa

Du Tran[2] trandu@fb.com

[1]King Abdullah University of Science and Technology (KAUST) [2]Facebook AI

http://humamalwassel.com/publication/xdc

[Alwassel et al., Self-Supervised Learning by Cross-Modal Audio-Video Clustering. Neural Information Processing Systems (NeurIPS), 2020
http://www.humamalwassel.com/publication/xdc/]



Learning: Alternate cross-modal clustering and classsifier learning.

Training data: Instagram video dataset (IG65M)

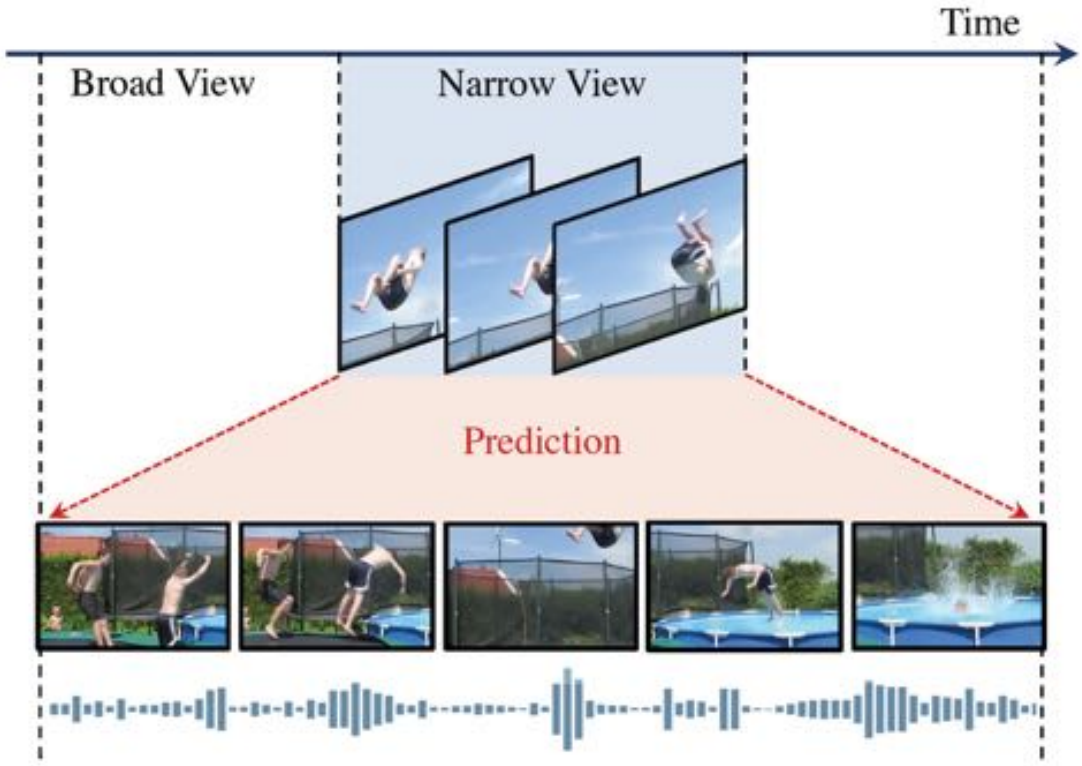
Evaluation: action recognition on HDMB and UCF101 datasets

Result: state-of-the-art performance from AV only.

Audio-video regression

Broaden Your Views for Self-Supervised Video Learning

Adrià Recasens[1] Pauline Luc[1] Jean-Baptiste Alayrac[1] Luyu Wang[1]
Florian Strub[1] Corentin Tallec[1] Mateusz Malinowski[1] Viorica Pătrăucean[1] Florent Altché[1]
Michal Valko[1] Jean-Bastien Grill[1] Aäron van den Oord[1] Andrew Zisserman[1,2]
[1]DeepMind [2]VGG, Dept. of Engineering Science, University of Oxford



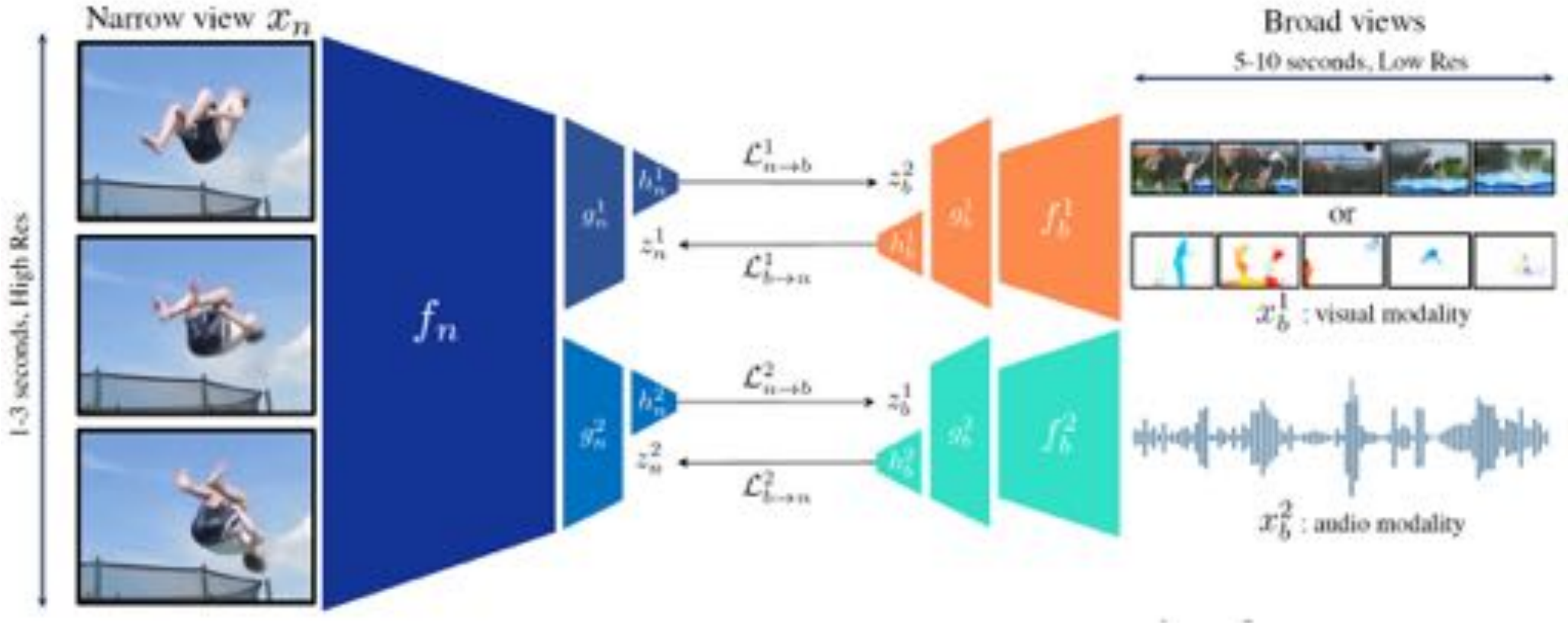
[Recasens et al., Broaden Your Views for Self-Supervised Video Learning, 2021 https://arxiv.org/abs/2103.16559]

Audio-video regression

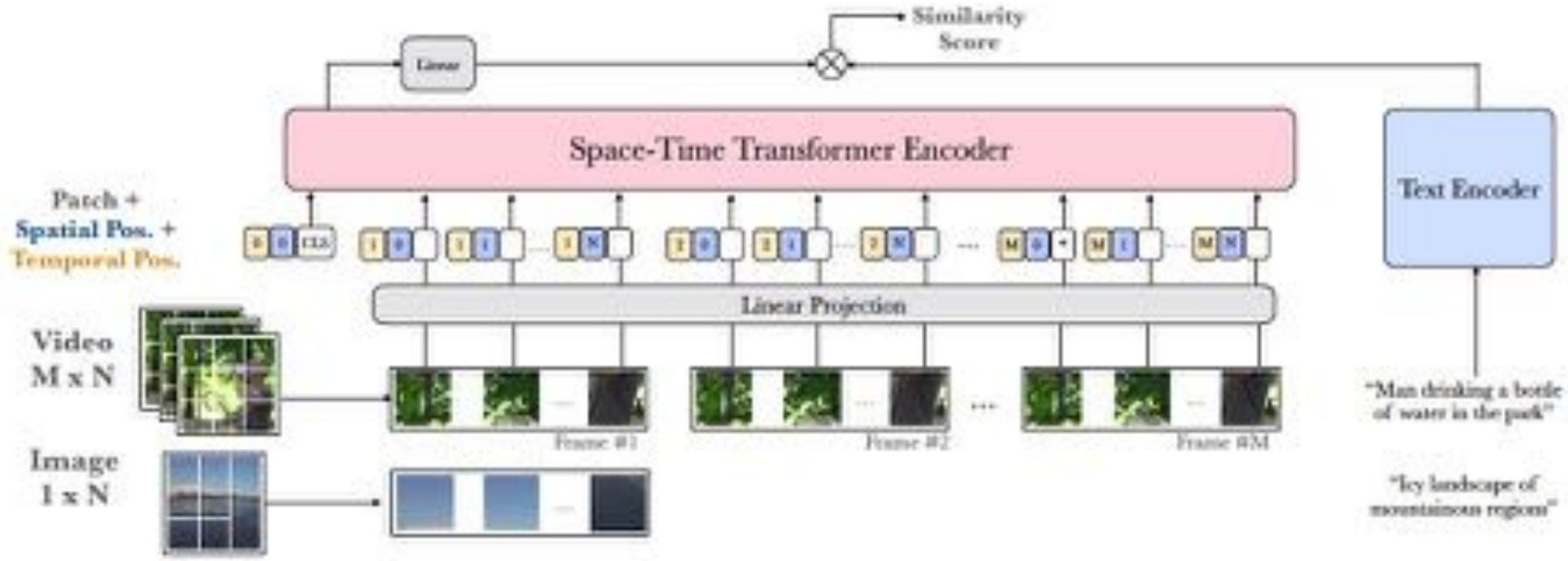
Training: Audio and video (Audioset dataset, 1.9M videos)

Evaluation: Action recognition (HMDB, UCF101, Kinetics)

Result: State-of-the-art on self-supervised benchmarks



Combine video with images



[Bain et al., Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval, https://arxiv.org/abs/2104.00650]

Other examples of meta-data: location and time

Google Street View: Time Machine

Example: Learn representation for place recognition

Answer “Where am I?” using street-view time-machine imagery

Has to deal with viewpoint, illumination, changes over time, …

Query



Match



[Arandjelovic et al., CVPR 2016]

Example: Visual localization in changing conditions

[Sattler et al., CVPR 2018]

https://www.visuallocalization.net/

Beyond classification: Image generation and editing

StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery, Patashnik et al.
https://arxiv.org/abs/2103.17249

See also: Paint by word, Bau et al., https://arxiv.org/abs/2103.10951

Learning from video: beyond action recognition

Actions often modify **states of object**.

Empty cup → **Fill** → **Full** cup

State 1 → **Action** → *State 2*

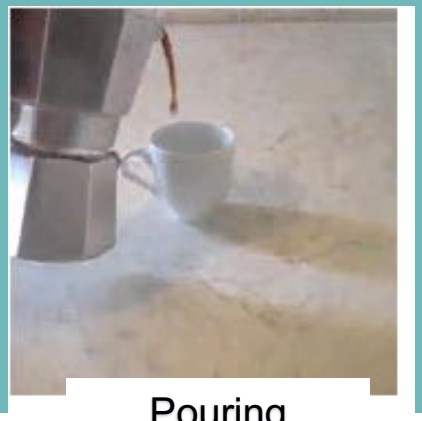
Also, e.g. **open** a *door,* **fill** a *water bottle,* ***cut*** *bread,…*

Can we learn the set of **actions** and **object states** from data?

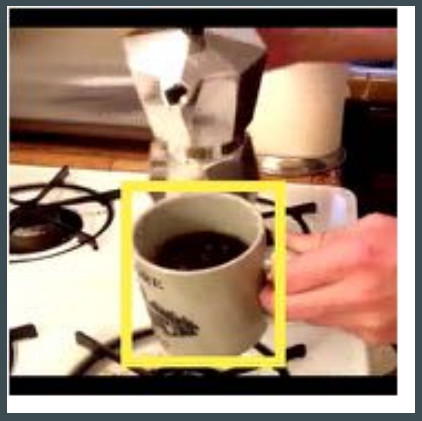
[Alayrac et al., ICCV 2017]

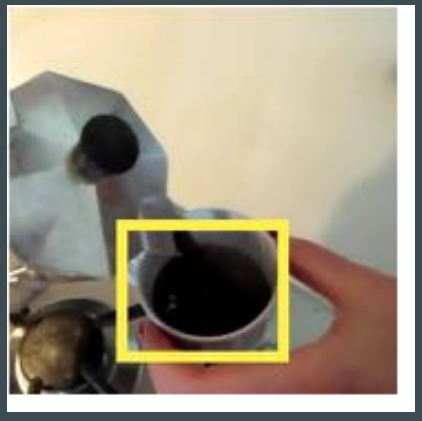
Pour coffee

Empty cup



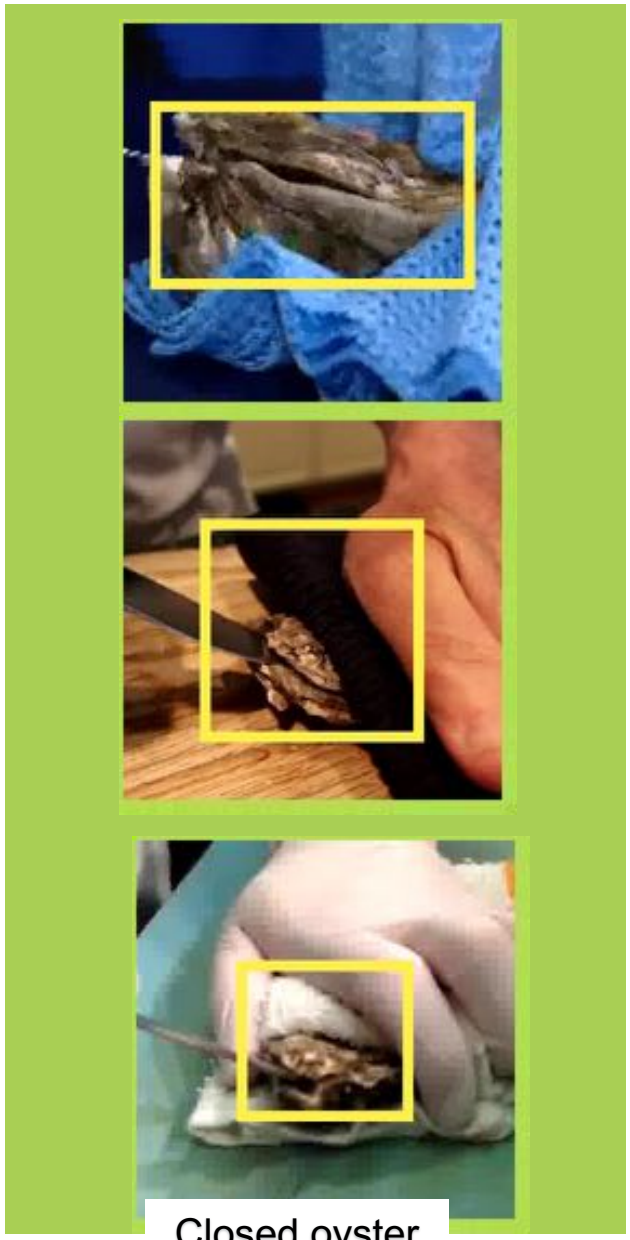
Pouring





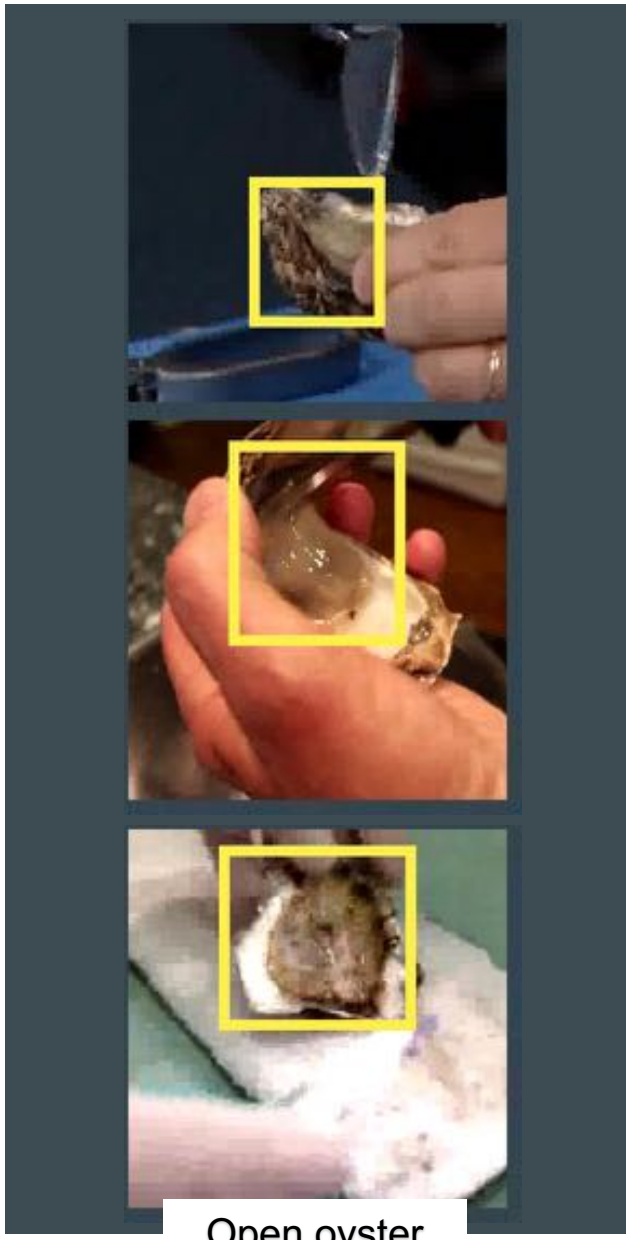
Full cup

Open oyster



Closed oyster

Open

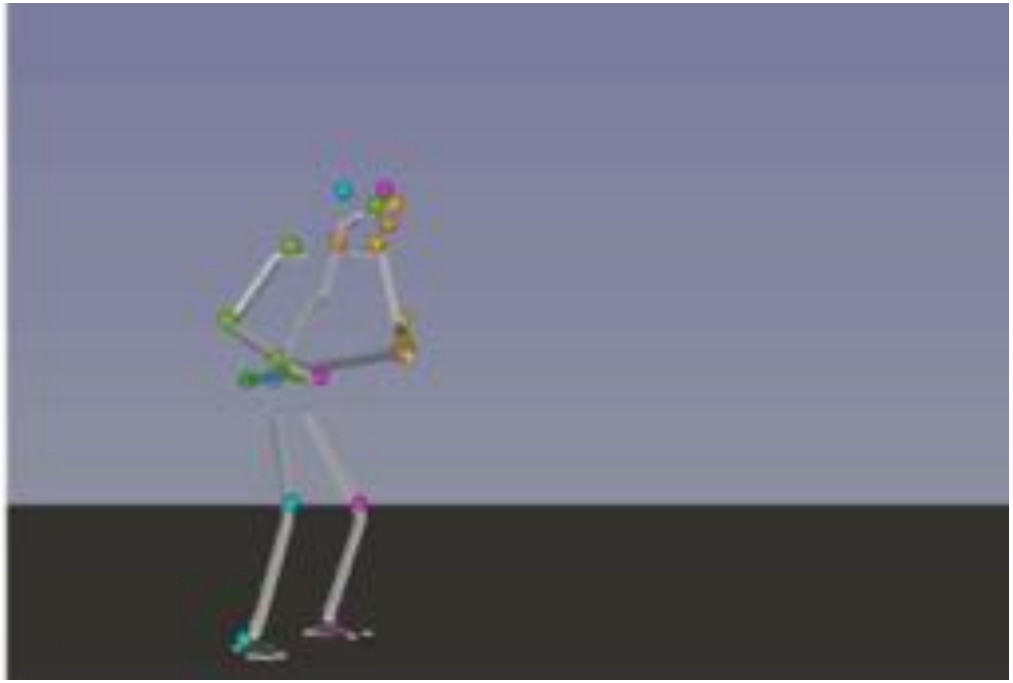


Open oyster

Learning motion from video

Input:
- a monocular RGB video

Output:
- Person & object 3D motion trajectories
- Contact positions and contact forces



[Li, Sedlar, Carpentier, Mansard, Laptev, Sivic, CVPR 2019]

Video results

Estimating 3D Motion and Forces of
Person-Object Interactions from Monocular Video

Supplementary video results

Learning to Use Tools by Watching Videos

Input: instructional video from YouTube

Output: tool manipulation skill transferred to a robot

[K. Zoryna et al., 2021]

Summary I.: learning from video, language, audio

Training datasets:

- HowTo100M (15 years), Video, Audio, Transcribed narration
- Audioset (~1 year) - Video, Audio
- YouTube8M (13 years) - Video, Audio, Tags
- InstaGram65M (21 years) - non-public, Video, Audio, Tags
- Kinetics 600 (0.1 year) - Video, Audio, labelled, human action classes

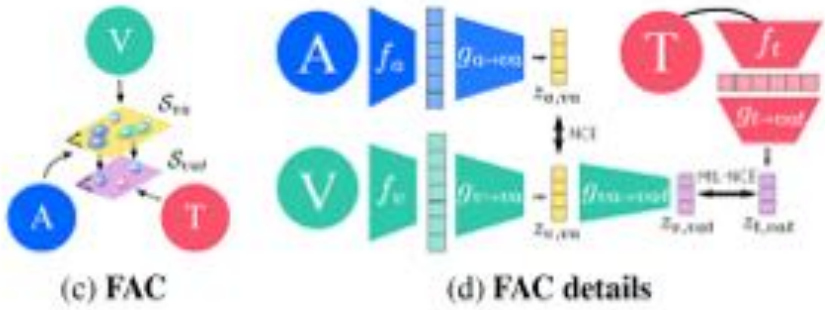
Downstream datasets and tasks:

- Action recognition (HDMB, UCF101, Kinetics)
- Action localization (COIN, YouTube8M Segments, CrossTask)
- Audio recognition (AudioSet, Environment Sound Dataset)
- Text to video retrieval (MSR-VTT, YouCook2, LSMDC)
- VideoQA (iVQA, MSRVTT-QA, MSVD-QA, ActivityNet-QA, How2QA)

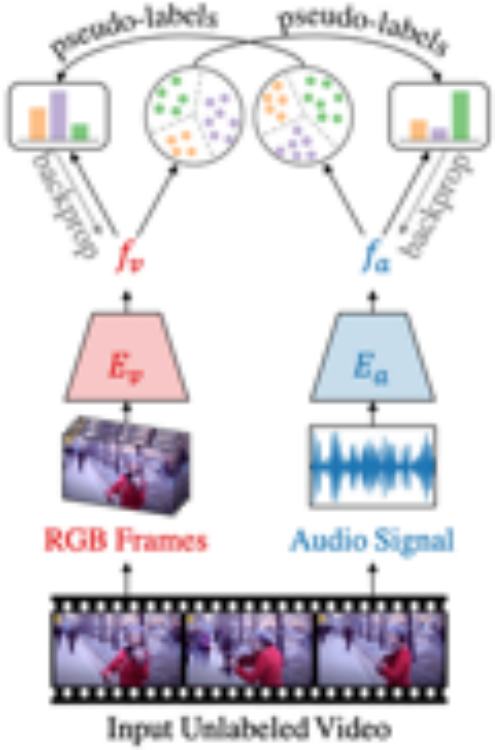
Multi-modal self-supervised learning is approaching or in some cases surpassing supervised learning.

Summary II.: main classes of methods

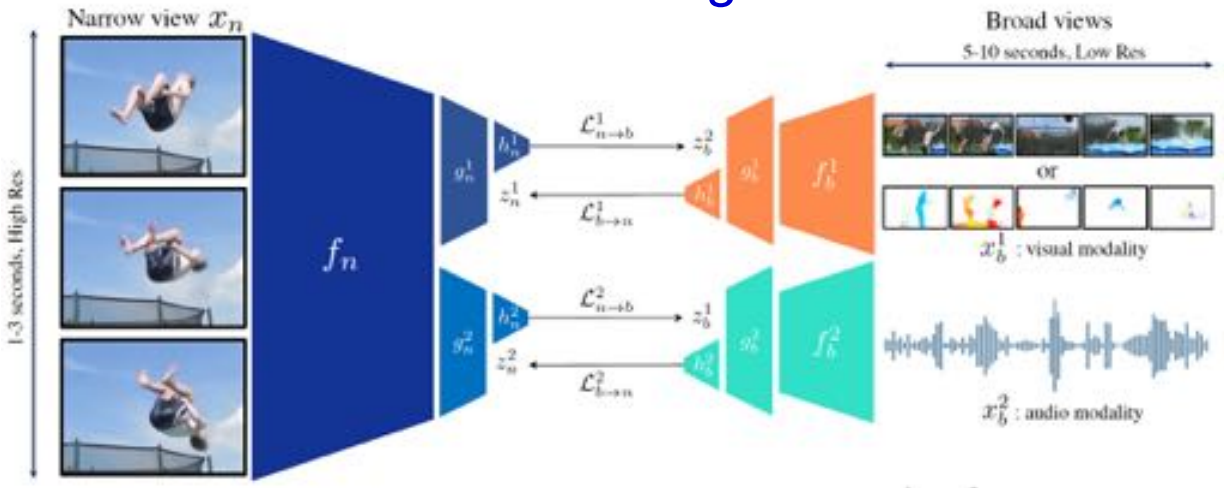
1. Multi-modal contrastive learning



2. Cross-modal clustering



3. Cross-modal regression



Summary III.: Modalities and outlook

Modalities:

- Video (appearance and motion), Language and Audio
- Still images

Towards learning embodied agents:

- Learn visuomotor representations with the help of video
- Additional sensors:
  - force, touch and contact,
  - Depth (Lidar, Stereo, …)
  - Odometry/localization/Inertial Measurement Units (ego motion)
- Interaction with the environment (reinforcement learning)

References I.

1. Multi-modal contrastive learning:

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, https://arxiv.org/abs/1912.06430.

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, Andrew Zisserman, Self-Supervised MultiModal Versatile Networks, NeurIPS 2020, https://arxiv.org/abs/2006.16228.

See also self-supervised contrastive learning for still images:

SimCLR: Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton, A Simple Framework for Contrastive Learning of Visual Representations, https://arxiv.org/abs/2002.05709, 2020.

MoCo: Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick, Momentum Contrast for Unsupervised Visual Representation Learning, 2019, https://arxiv.org/abs/1911.05722

Xinlei Chen, Haoqi Fan, Ross Girshick, Kaiming He, Improved Baselines with Momentum Contrastive Learning, 2020. https://arxiv.org/abs/2003.04297

Recent very large scale image-language models:

Open-AI CLIP: Radford et al., Learning Transferable Visual Models From Natural Language Supervision, 2021, https://arxiv.org/abs/2103.00020

Google ALIGN: Jia et al., Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, 2021. https://arxiv.org/abs/2102.05918

Origins of joint visual-language embeddings :

Jason Weston, Samy Bengio, Nicolas Usunier, Wsabie: Scaling up to large vocabulary image annotation, AAAI 2011.

A Frome, G Corrado, J Shlens, S Bengio, J Dean, MA Ranzato, T Mikolov, Devise: A deep visual-semantic embedding model, NIPS 2013

T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean, Distributed representations of words and phrases and their compositionality, Neural information processing systems, 2013.

Y Gong, Q Ke, M Isard, S Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, International journal of computer vision 106 (2), 210-233, 2014. https://arxiv.org/pdf/1212.4522.pdf

References II.

2. Multi-modal clustering:

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, Du Tran, Self-Supervised Learning by Cross-Modal Audio-Video Clustering, NeurIPS 2020. http://www.humamalwassel.com/publication/xdc/

See also self-supervised clustering for still images:

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In ECCV, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In NeurIPS, 2020.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In ICLR, 2020

3. Regression methods:

Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Patraucean, Florent Altché, Michal Valko, Jean-Bastien Grill, Aäron van den Oord, Andrew Zisserman, Broaden Your Views for Self-Supervised Video Learning. 2021, https://arxiv.org/abs/2103.16559

Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Perez, and Matthieu Cord. Learning representations by predicting bags of visual words. In CVPR, 2020

Pierre H Richemond, Jean-Bastien Grill, Florent Altche Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, et al. Byol works even without batch statistics. In NeurIPS (SSL Workshop), 2020

Jean-Bastien Grill, Florian Strub, Florent Altch´e, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In NeurIPS, 2020.

Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, Patrick Perez, OBoW: Online Bag-of-Visual-Words Generation for Self-Supervised Learning, CVPR 2021, https://arxiv.org/pdf/2012.11552.pdf

See also an excellent tutorial on self-supervised learning at CVPR 2021:

https://gidariss.github.io/self-supervised-learning-cvpr2021/

References III.

4. (Non-exhaustive) list of the downstream tasks and datasets

HowTo100M: Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev and Josef Sivic, HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, ICCV 2019, https://arxiv.org/abs/1906.03327

RareAct: Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic and Andrew Zisserman, RareAct: A video dataset of unusual interactions, https://arxiv.org/abs/2008.01018, 2020.

VQA: Antoine Yang, Antoine Miech, Ivan Laptev, Josef Sivic and Cordelia Schmid, Just Ask: Learning to Answer Questions from Millions of Narrated Videos, https://arxiv.org/abs/2012.00451, 2021.

Action recognition:

UCF101: https://www.crcv.ucf.edu/data/UCF101.php

HMDB-51: https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

Kinetics: https://deepmind.com/research/open-source/kinetics

Action localization:

CrossTask: https://github.com/DmZhukov/CrossTask

COIN: https://coin-dataset.github.io/

YouTube8M: http://research.google.com/youtube8m/

Audio classification

AudioSet: http://research.google.com/audioset/

ESC: https://github.com/karolpiczak/ESC-50

Text-to-video retrieval

MSRVTT: https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/

YouCook2: http://youcook2.eecs.umich.edu/

LSMDC: https://sites.google.com/site/describingmovies

References IV.

5. Other modalities and tasks

Vision and touch for robotics:

Michelle A. Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, Jeannette Bohg, Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks, ICRA 2019. Best paper award. https://ai.stanford.edu/blog/selfsupervised-multimodal/

Estimating motion and forces of tool manipulation from Youtube videos

Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard and Josef Sivic. Estimating 3D Motion and Forces of Person-Object Interactions from Monocular Video, CVPR 2019.

Towards learning reward signals for reinforcement learning in robotics

Jean-Baptiste Alayrac, Josef Sivic, Ivan Laptev and Simon Lacoste-Julien, Joint Discovery of Object States and Manipulation Actions In Proc. ICCV 2017. https://www.di.ens.fr/willow/research/objectstates/

Learning visual representations from geotagged datasets:

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, Josef Sivic, NetVLAD: CNN architecture for weakly supervised place recognition, CVPR 2016, https://arxiv.org/abs/1511.07247

Combining still images, videos and language:

Max Bain, Arsha Nagrani, Gul Varol, Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval, 2021. https://arxiv.org/pdf/2104.00650.pdf

Generating and editing images using natural language:

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, Dani Lischinski, Styleclip: Text-driven manipulation of stylegan imagery, 2021, https://arxiv.org/abs/2103.17249

David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, Antonio Torralba, Paint by Word, 2021. https://arxiv.org/abs/2103.10951

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever, Zero-Shot Text-to-Image Generation, 2021, https://arxiv.org/abs/2102.12092

Thank you