Weakly-supervised and self-supervised visual recognition

Ivan Laptev

Ivan.Laptev@mbzuai.ac.ae
https://www.di.ens.fr/~laptev
Visiting professor, MBZUAI, United Arab Emirates
External member, Willow Team, Inria, DI ENS, Paris

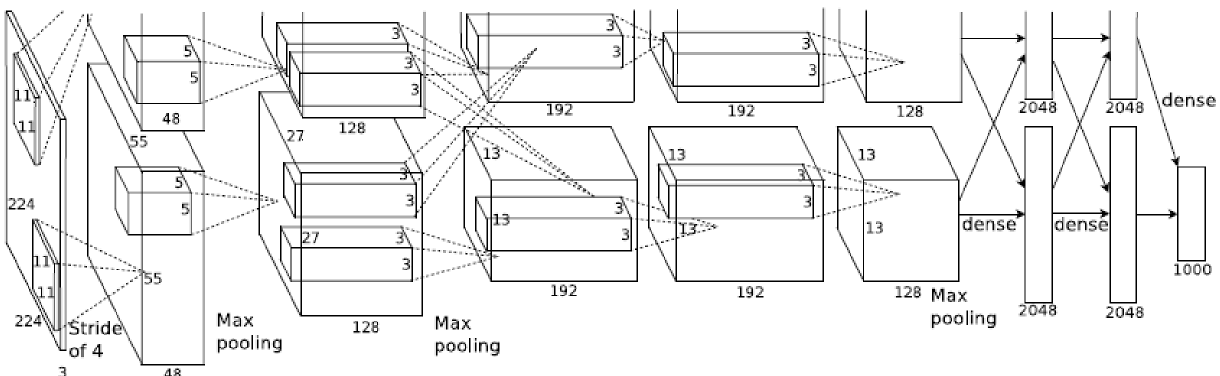
informatics mathematics
ENS
1794
MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

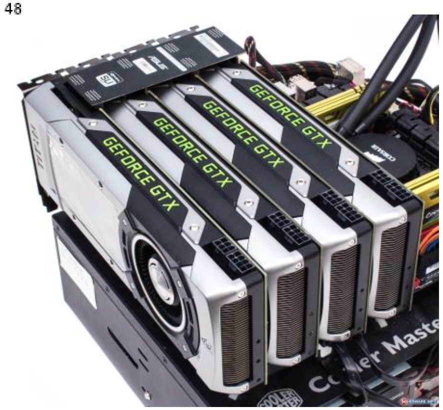
Announcements

- Final Project Topics are out
- Final Project proposals are due Dec 5
  - Submit 1-page PDF project proposal
  - Add “Topic A” … “Topic X” to the title
  - Check you can claim Google Credits (wait for email)
  - Whenever possible, aim to reproduce published results using comparable experimental settings.
  - Use standard experimental settings so your results can be compared to others
  - Start working on the project, your will get feedback on your project proposal

Ingredients of modern vision methods

AlexNet [Krizhevsky et al. 2012]
~60M parameters

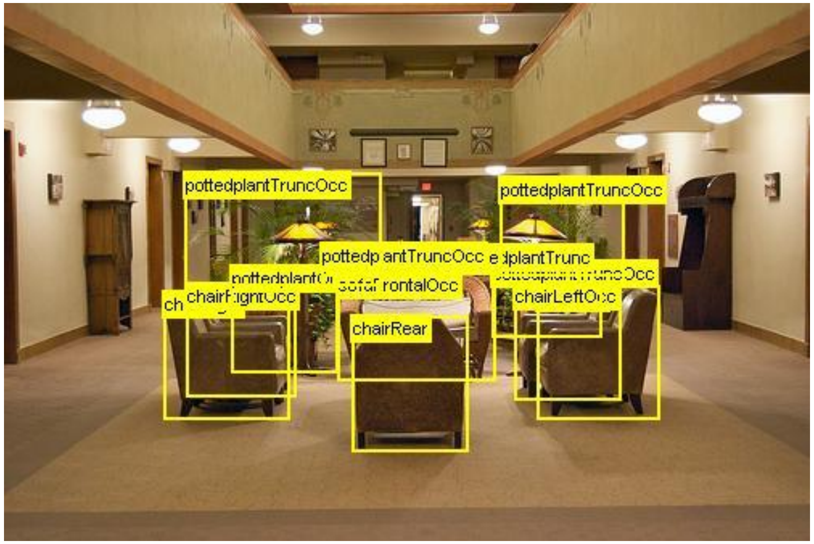




Manual supervision

Problems with manual supervision

- Expensive



- Ambiguous

Table? Dining table? Desk? …

This lecture:

How to avoid manual supervision?

Part I:
Weakly-supervised learning

Coarse or cheap labels

Part II:
Self-supervised learning

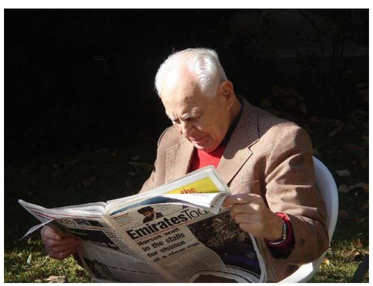
No labels

Preview:

Weakly-supervised learning

Coarse or cheap labels

image-level labels:



+ ✓ Person
✓ Chair
✗ Airplane

Self-supervised learning

No labels

Pre-trained visual representation

Can we train object detection without bounding box annotations?

Image-level labels: Bicycle, Person

Motivation: image-level labels are plentiful

“Beautiful red leaves in a back street of Freiburg”

[Kuznetsova et al., ACL 2013]
http://www.cs.stonybrook.edu/~pkuznetsova/imgcaption/captions1K.html

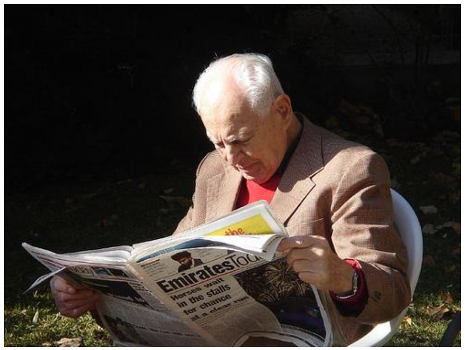
Motivation: image-level labels are plentiful

"Public bikes in Warsaw during night"

https://www.flickr.com/photos/jacek_kadaj/8776008002/in/photostream/

Goal

Training input



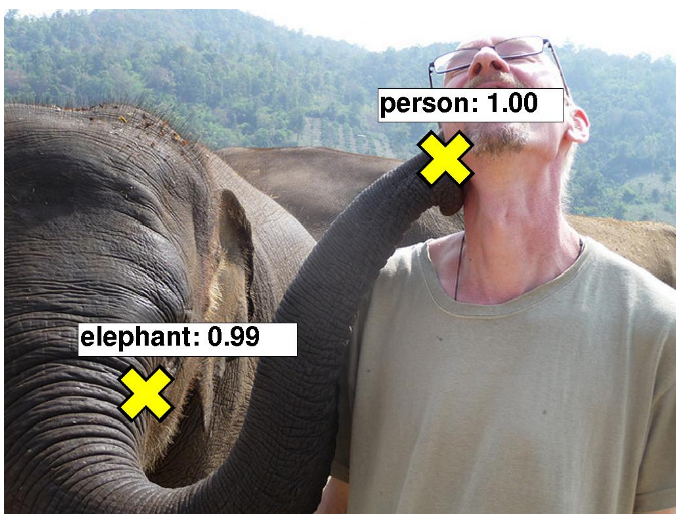
+

image-level labels:

- ✓ Person
- ✓ Chair
- ✗ Airplane
- …

- ✓ Reading
- ✗ Riding bike
- ✗ Running
- …

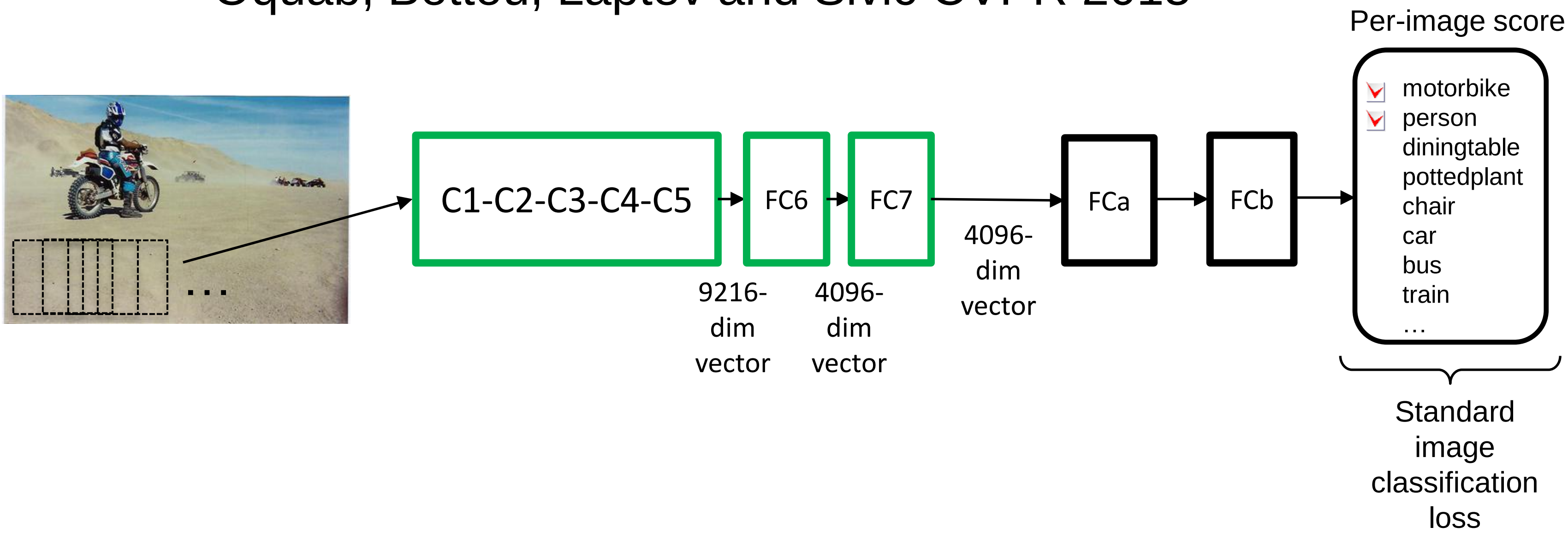
Test output



More details in http://www.di.ens.fr/willow/research/weakcnn/

Approach: search over object's location at the *training time*

Oquab, Bottou, Laptev and Sivic CVPR 2015



See also [Papandreou et al. '15, Sermanet et al. '14, Chaftield et al.'14]

Approach: search over object's location at the *training time*

Oquab, Bottou, Laptev and Sivic CVPR 2015

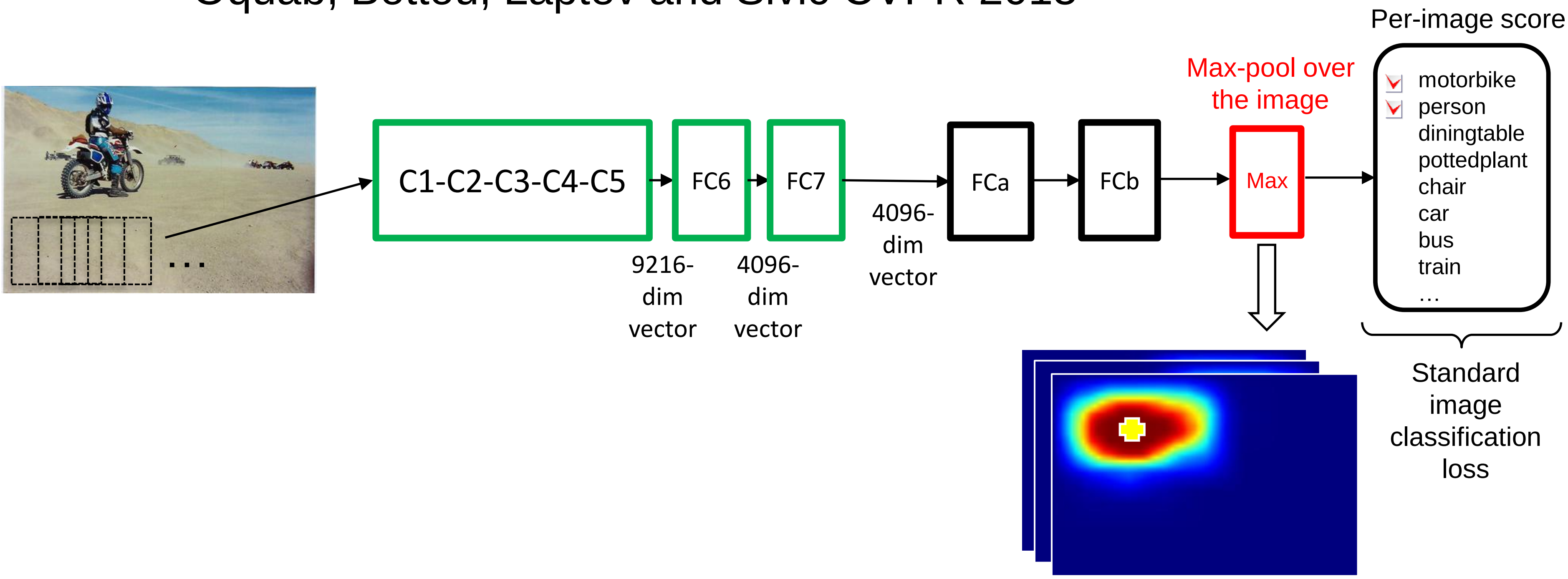
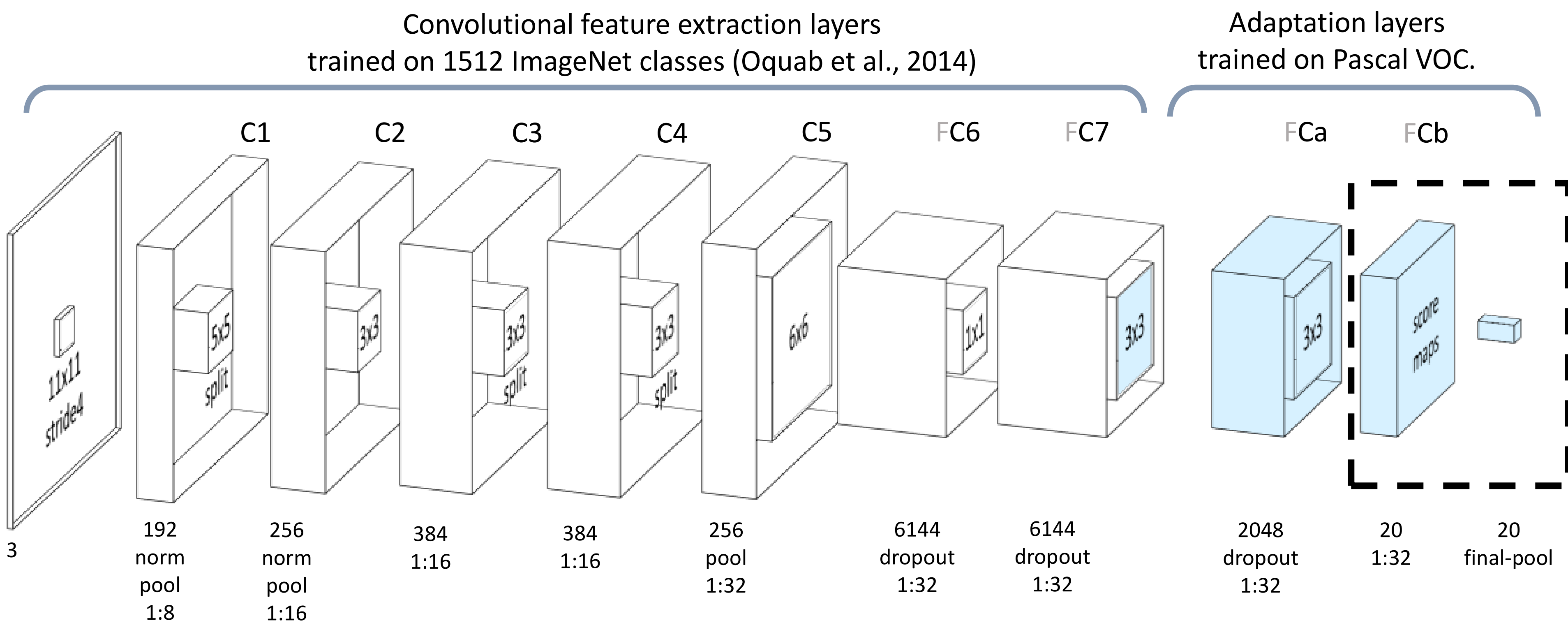
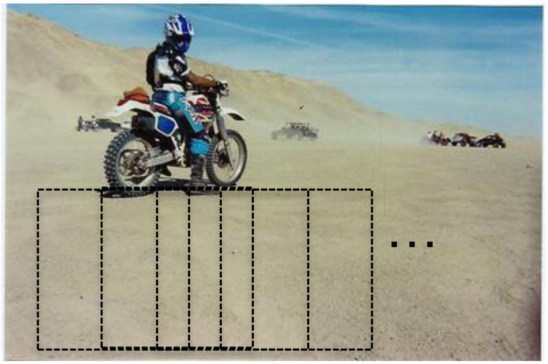


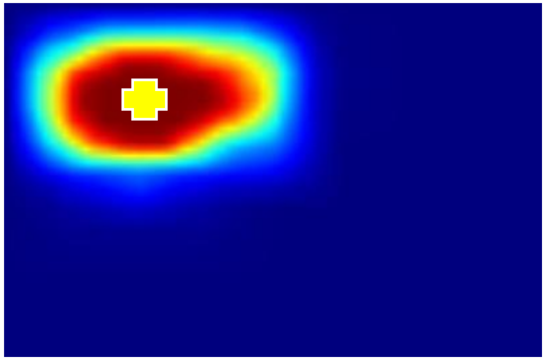
Image-level global max-pool per-class aggregation

See also [Papandreou et al. '15, Sermanet et al. '14, Chaftield et al.'14]

Image-level aggregation using global max-pool







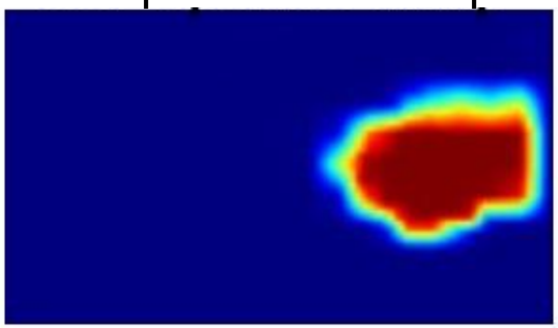
Training with global max-pooling

Training input:

\+ image-level labels:

- ✓ Airplane
- × Car
- × Chair …

Airplane score map



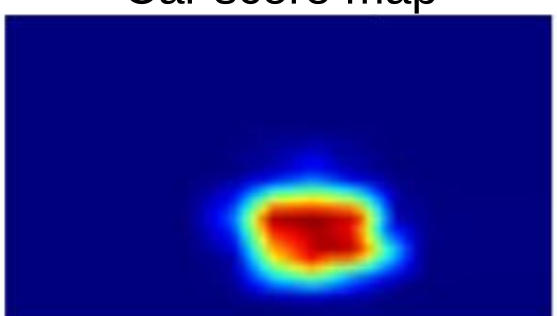
max-pool →

Correct label: increase score

⇩

Learn discriminative object parts

Car score map



max-pool →

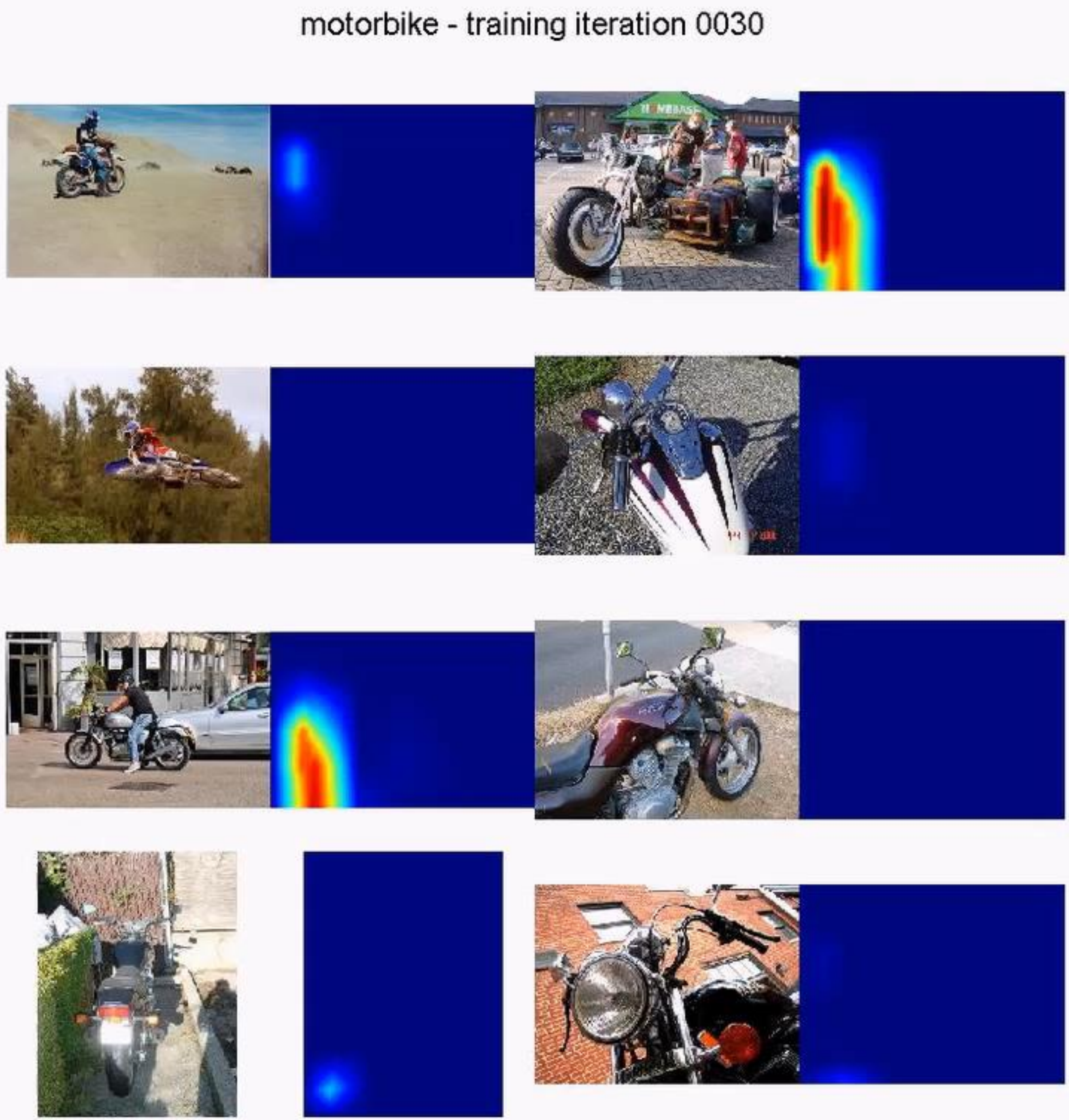
Incorrect label: decrease score

⇩

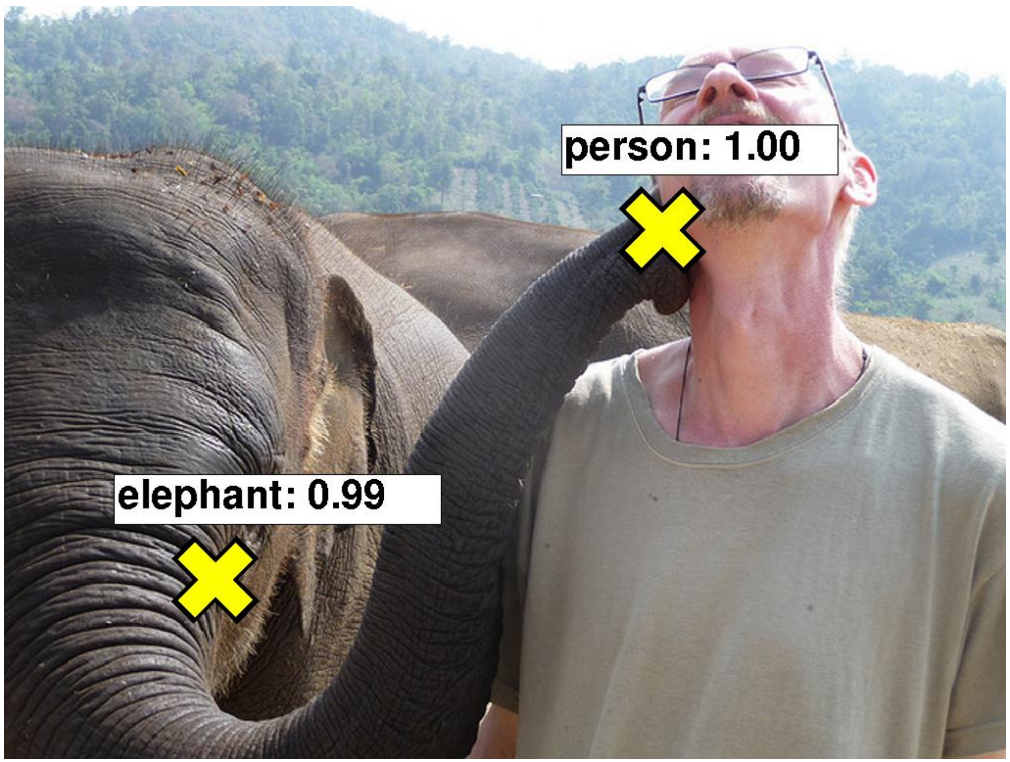
Suppress *Hard Negatives*

Training Motorbikes

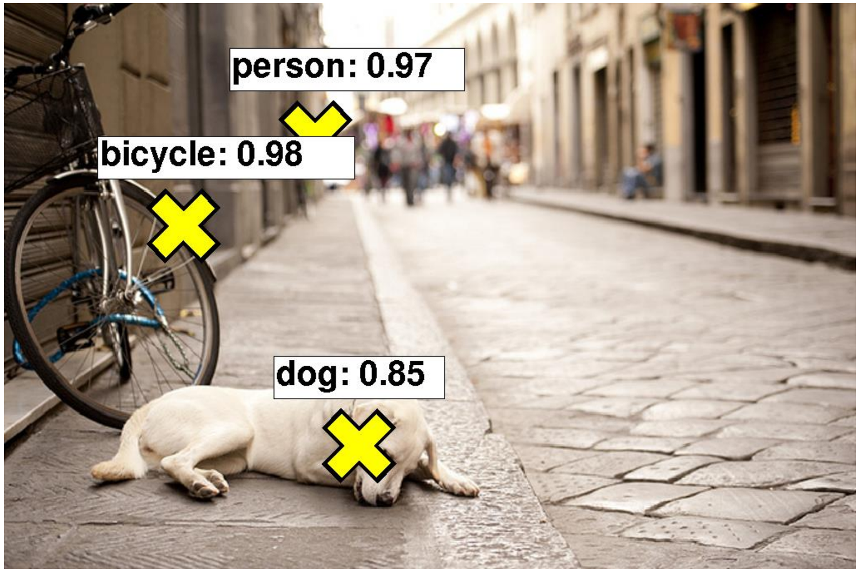
Evolution of localization score maps over training epochs

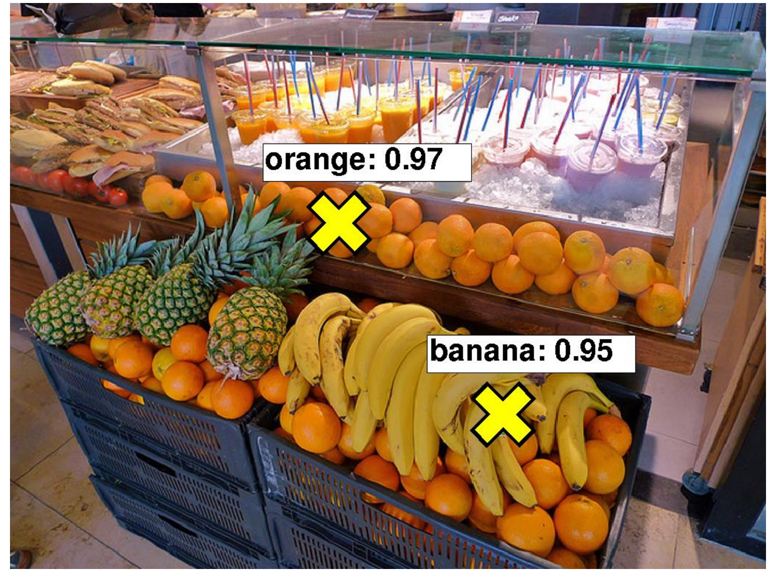


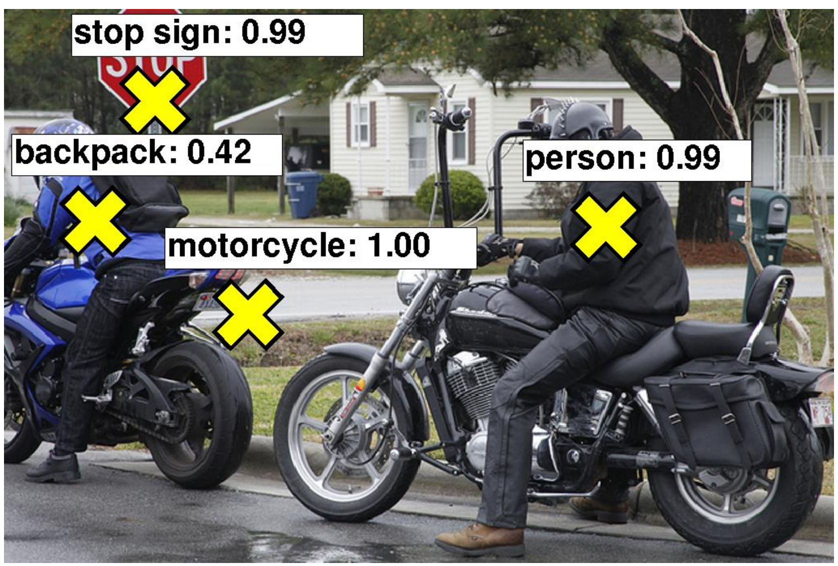
Test results on 80 classes in Microsoft COCO dataset



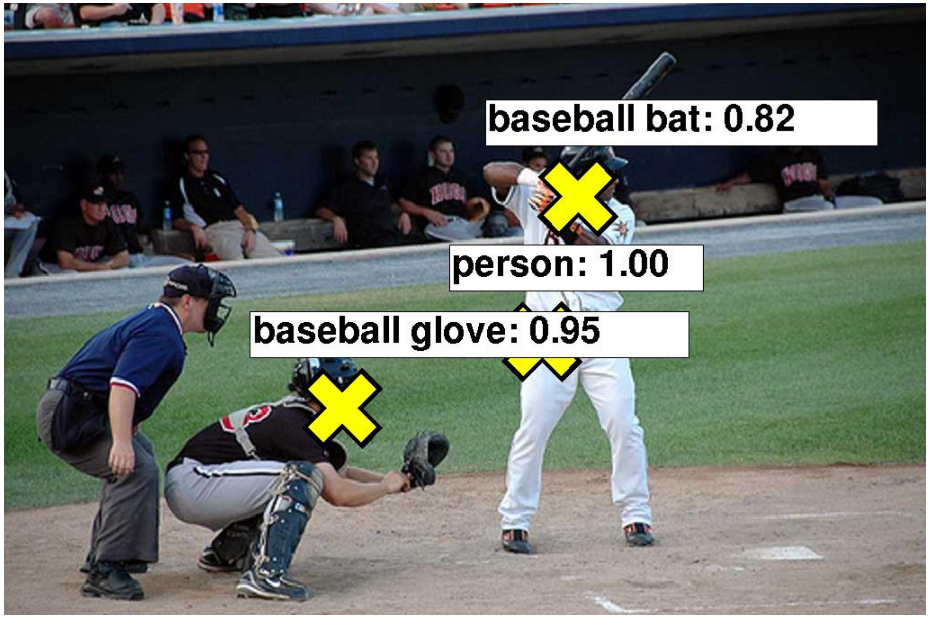
Test results on 80 classes in Microsoft COCO dataset

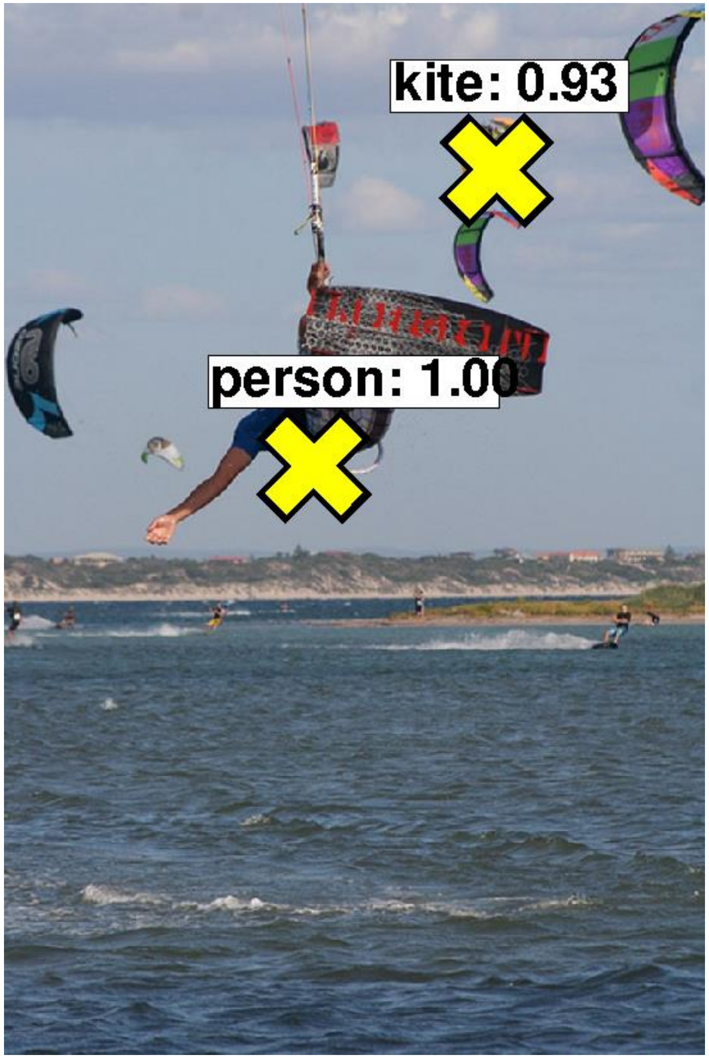


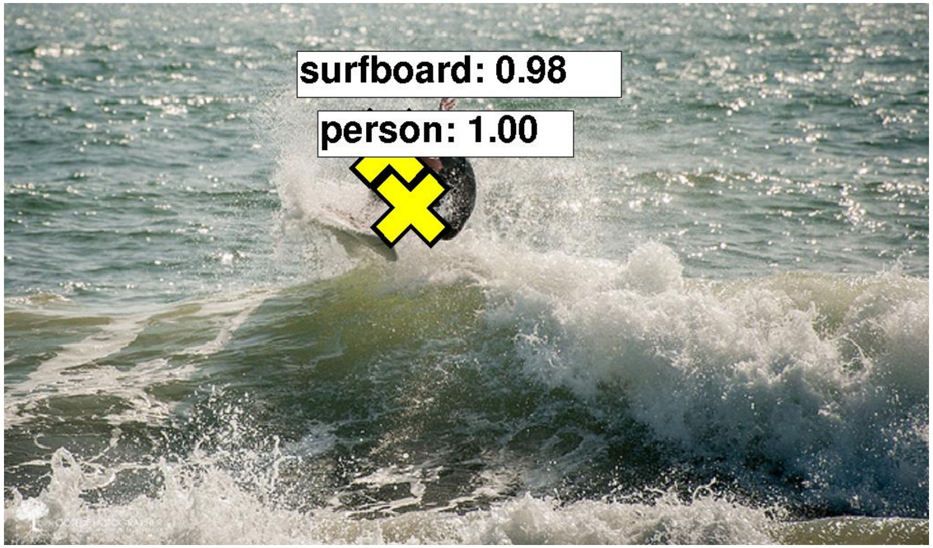




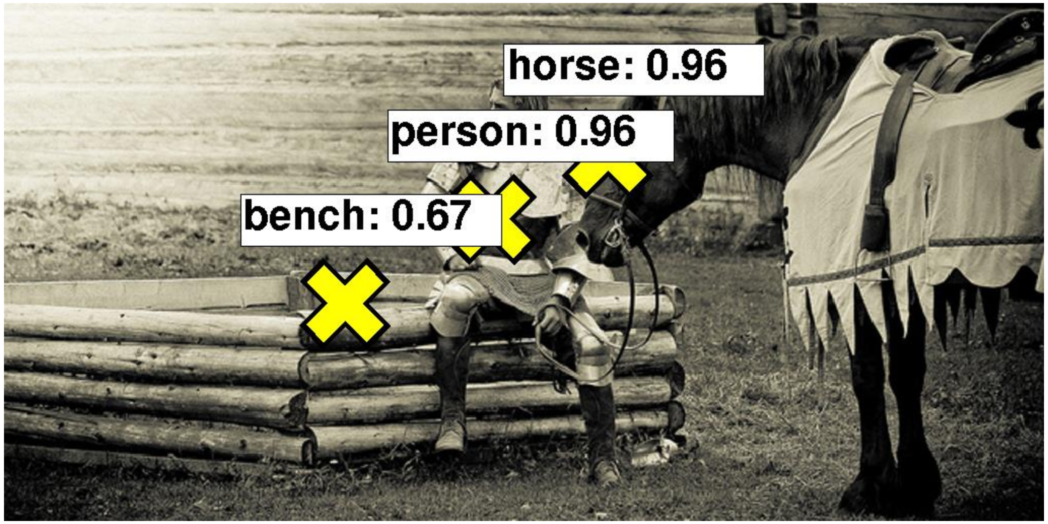
Test results on 80 classes in Microsoft COCO dataset

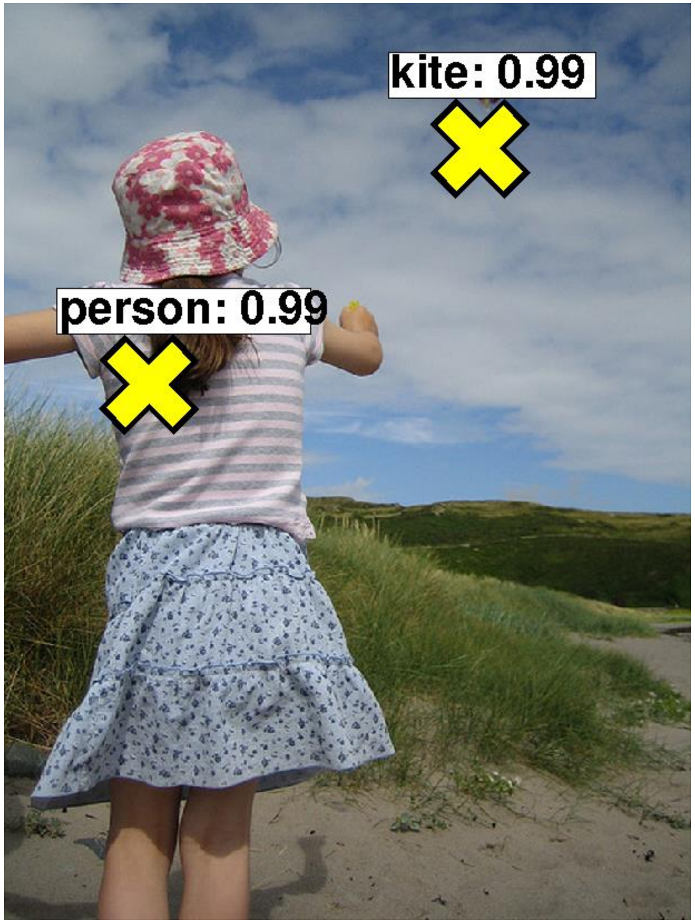


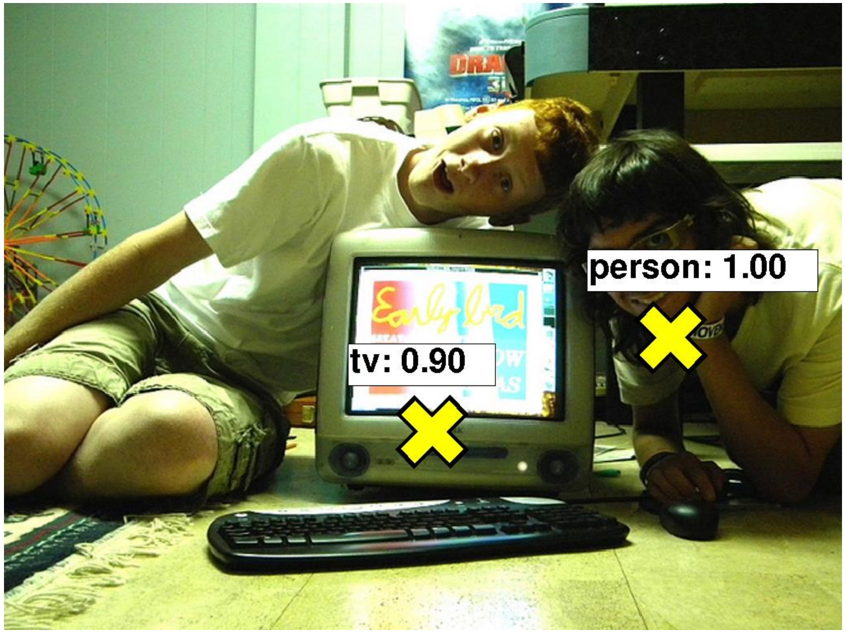




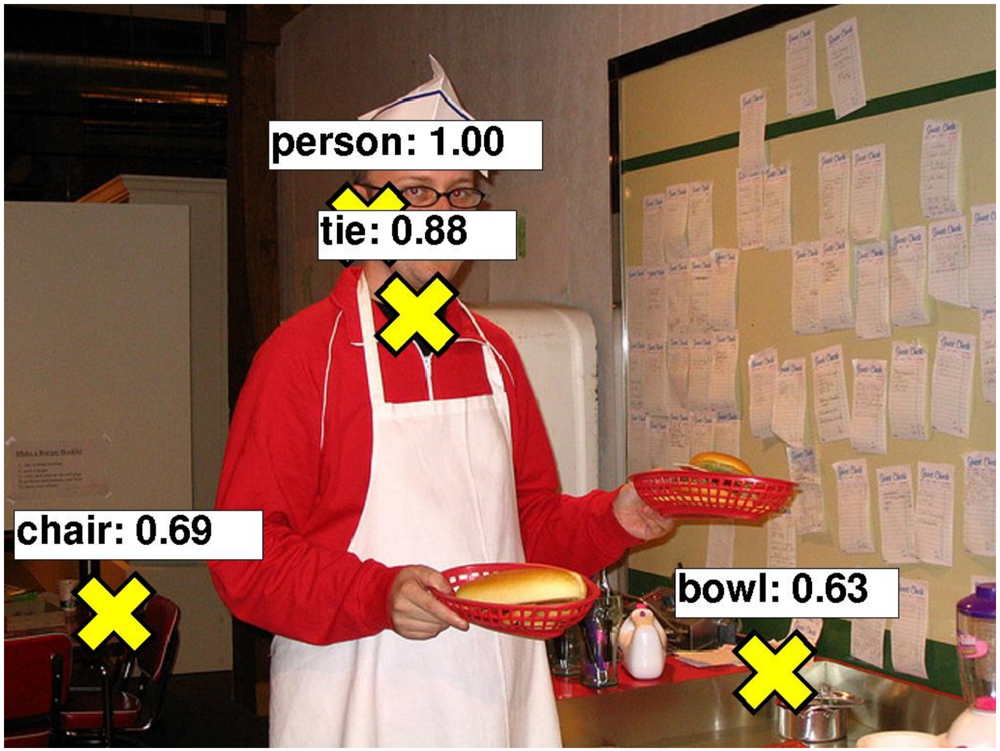
Test results on 80 classes in Microsoft COCO dataset

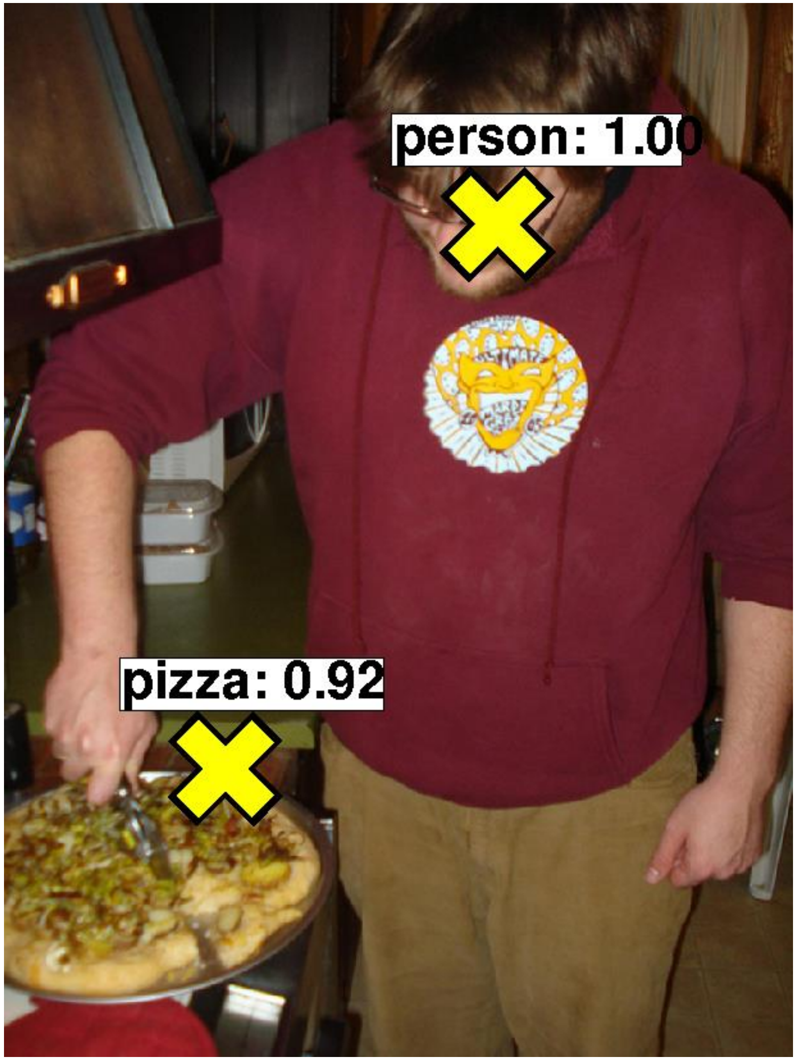






Test results on 80 classes in Microsoft COCO dataset

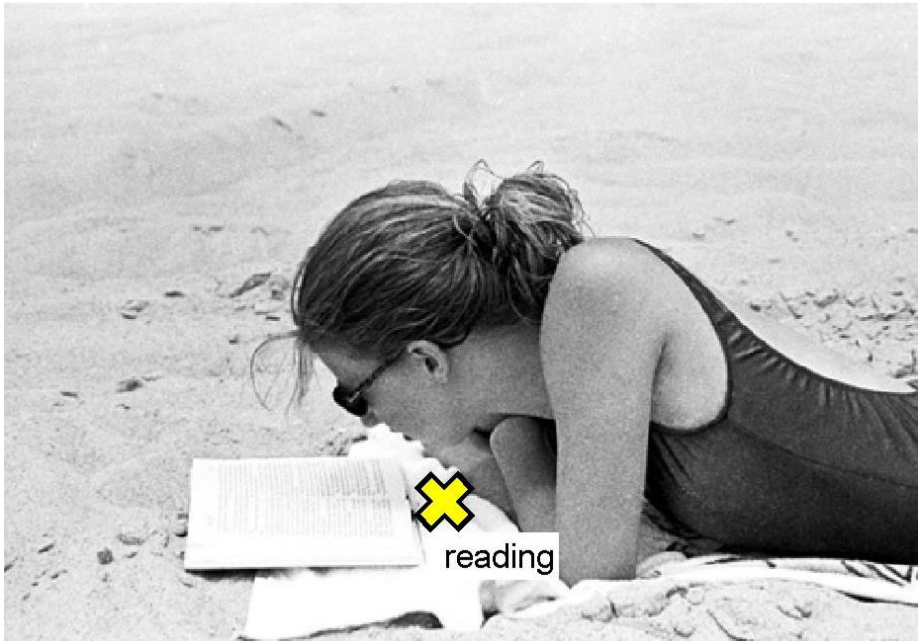


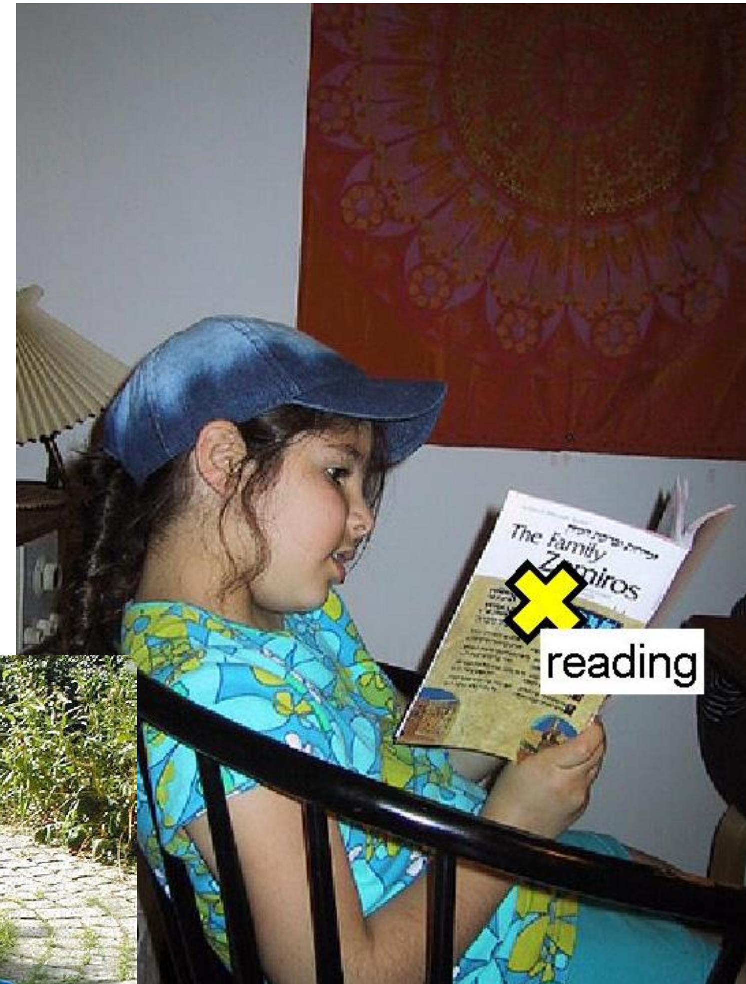


Results for weakly-supervised *action* recognition in Pascal VOC'12 dataset

Test results for **10 action classes** in Pascal VOC12

Test results for **10 action classes** in Pascal VOC12





Test results for **10 action classes** in Pascal VOC12

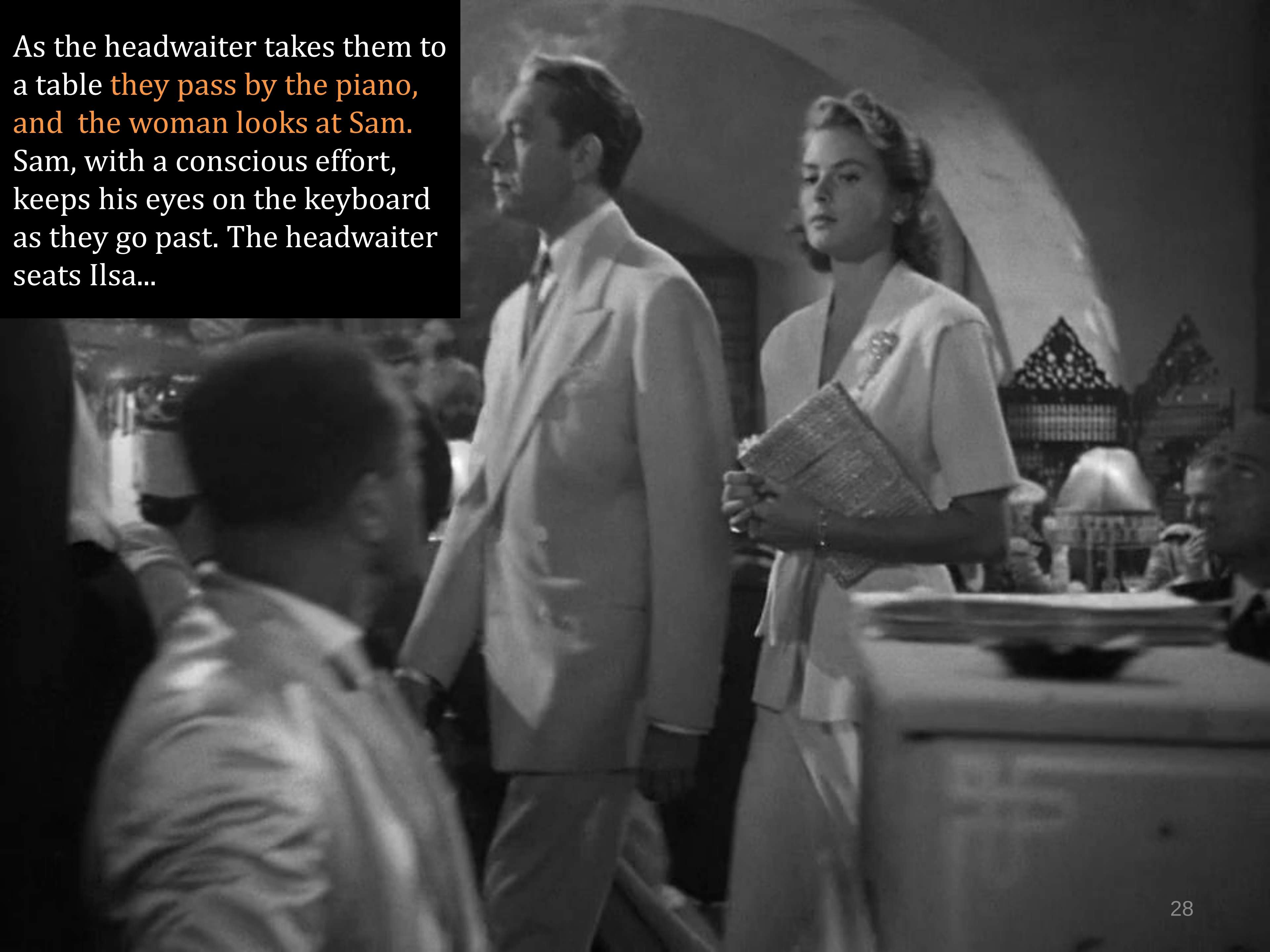
Test results for **10 action classes** in Pascal VOC12

Failure cases

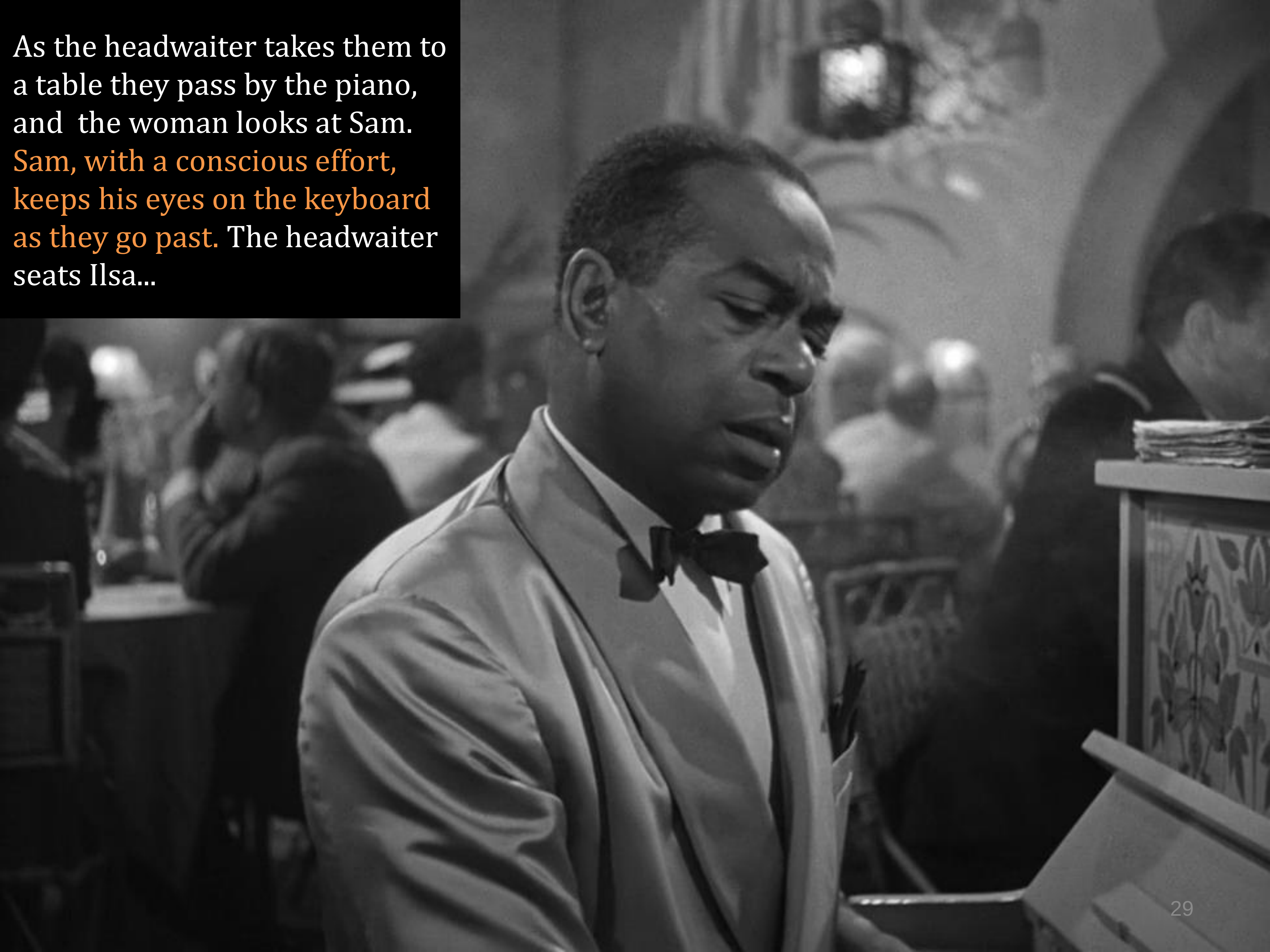
Weakly-supervised learning of actions *in video* from scripts and narrations

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

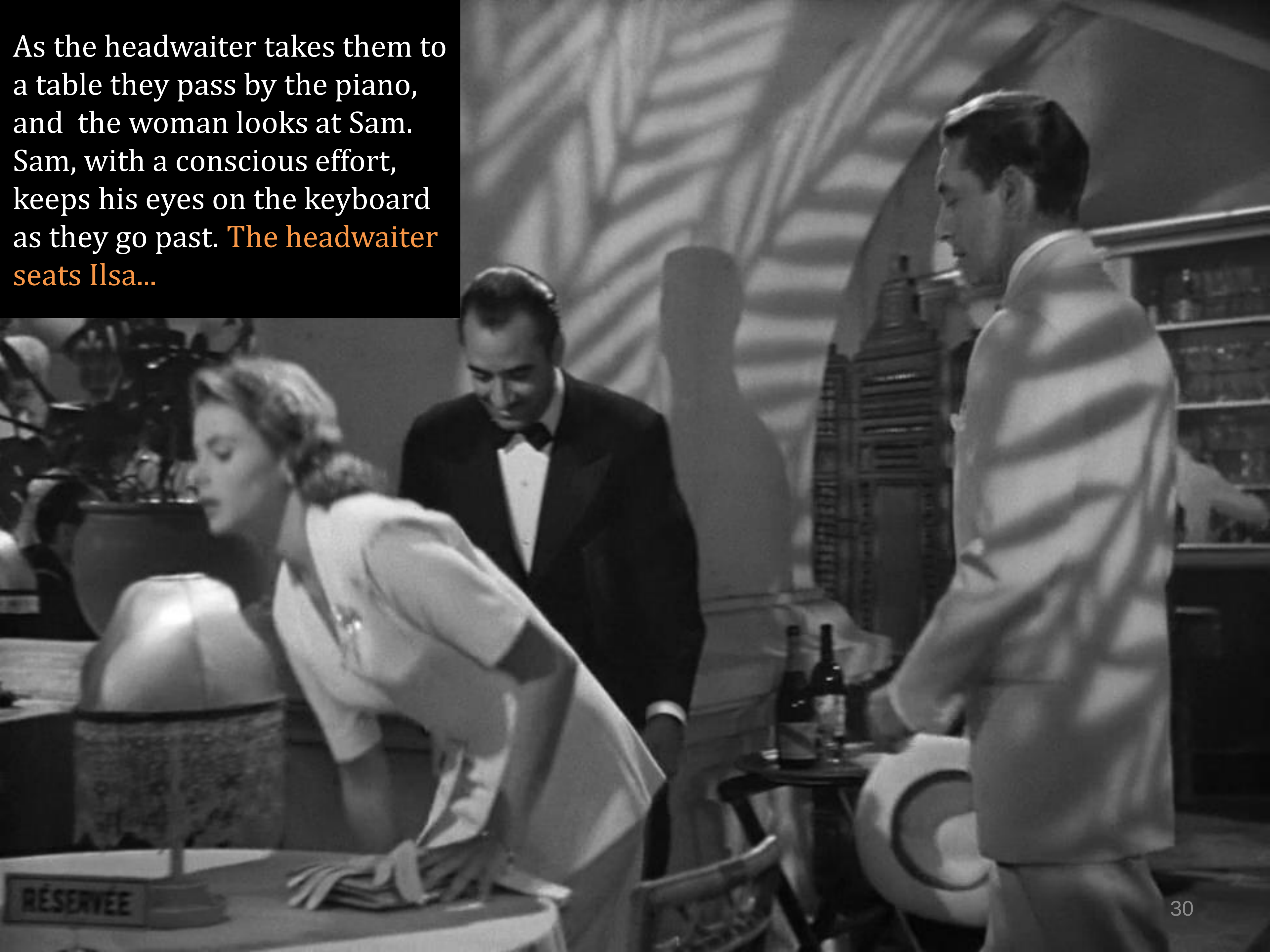
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

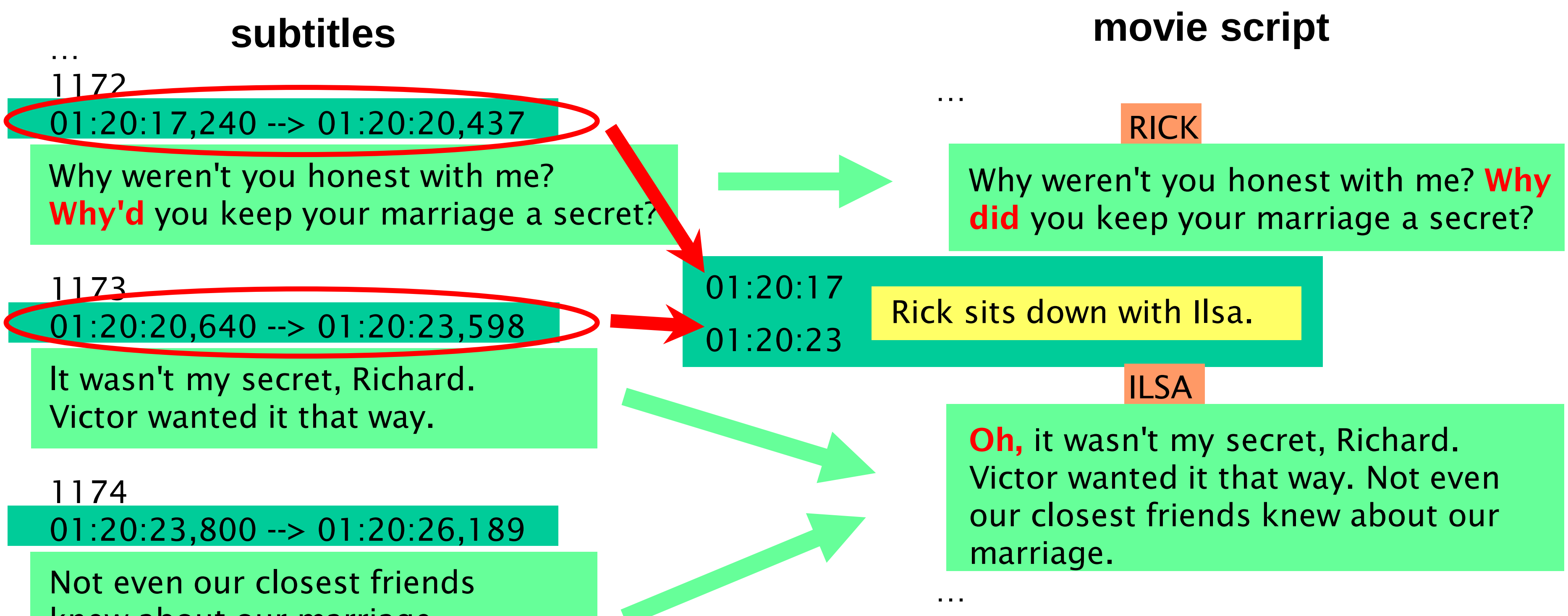


As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



Script-based video annotation

- Scripts available for >500 movies (no time synchronization)
  www.dailyscript.com, www.movie-page.com, www.weeklyscript.com …
- Subtitles (with time info.) are available for the most of movies
- Can transfer time to scripts by text alignment



[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Scripts as weak supervision

Challenges:

- Imprecise temporal localization
- No explicit spatial localization
- NLP problems, scripts ≠ training labels

"… Will gets out of the Chevrolet. …"
"… Erin exits her new truck…"

vs. Get-out-car

Subtitles

00:24:22 – 00:24:25
– Yes, Monsieur Laszlo. Right this way.

00:24:51 – 00:24:53
Two Cointreaux, please.

Script

Speech

Monsieur Laszlo. Right this way.

Scene description

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Speech

Two cointreaux, please.

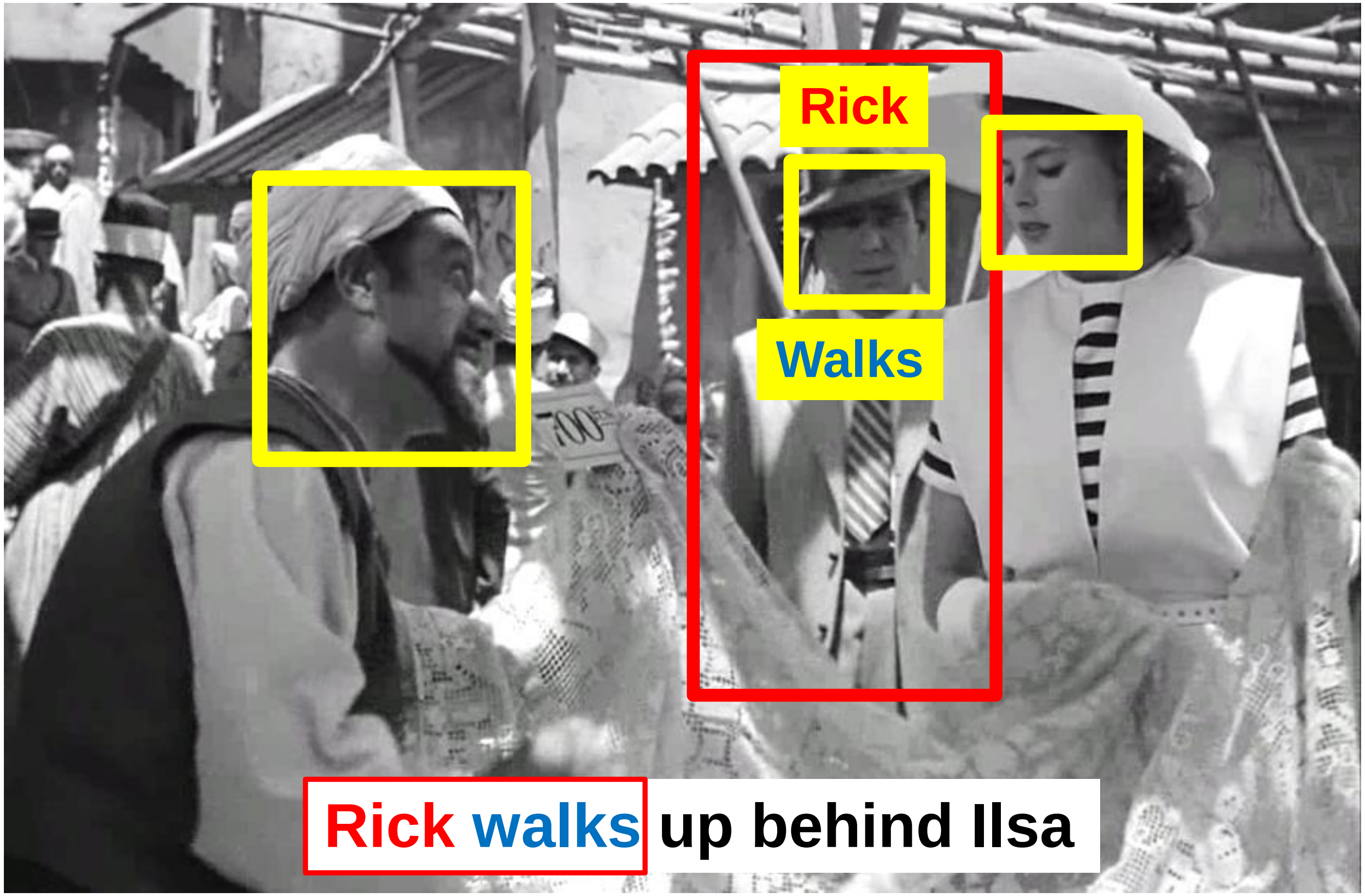
Joint Learning of Actors and Actions

[Bojanowski et al. ICCV 2013]

[Bojanowski, Bach, Laptev, Ponce, Schmid, Sivic, 2013]

Joint Learning of Actors and Actions

[Bojanowski et al. ICCV 2013]



[Bojanowski, Bach, Laptev, Ponce, Schmid, Sivic, 2013]

Formulation: Cost function

$$\frac{1}{N}\|Z - \phi(X)w - b\|_F^2 + \lambda_1\, Tr(w^T\, w)$$

Actor classifier

Actor labels

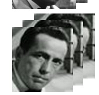
Actor image features

Rick
Ilsa
Sam

Formulation: Cost function

$$\frac{1}{N}\|Z - \phi(X)w - b\|_F^2 + \lambda_1 \, Tr(w^T \, w)$$

$$\begin{bmatrix} z_{11} & \cdots & z_{1p} & \cdots & z_{1P} \\ \vdots & & \vdots & & \vdots \\ z_{n_1 1} & \cdots & z_{n_1 p} & \cdots & z_{n_1 P} \\ z_{n_2 1} & \cdots & z_{n_2 p} & \cdots & z_{n_2 P} \\ z_{n_3 1} & \cdots & z_{n_3 p} & \cdots & z_{n_3 P} \\ \vdots & & \vdots & & \vdots \\ z_{N1} & \cdots & z_{Np} & \cdots & z_{NP} \end{bmatrix}$$


p = Rick

Weak supervision from scripts:

Person p appears at least once in clip N :

$$\sum_{n \in \mathcal{N}_i} z_{np} \geq 1$$

Formulation: Cost function

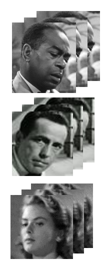
$$\frac{1}{N}\|Z-\phi(X)w-b\|_F^2+\lambda_1\ Tr(w^T\ w)$$

$$+\ \frac{1}{N}\|T-\psi(X)v-c\|_F^2+\lambda_2\ Tr(v^T\ v)$$

Weak supervision from scripts:

Action a appears at least once in clip N :

$$\sum_{n\in\mathcal{N}_i} t_{na}\geq 1$$



$$\begin{bmatrix} t_{11} & \cdots & t_{1a} & \cdots & t_{1A} \\ \vdots & & \vdots & & \vdots \\ t_{n_1 1} & \cdots & t_{n_1 a} & \cdots & t_{n_1 A} \\ t_{n_2 1} & \cdots & t_{n_2 a} & \cdots & t_{n_2 A} \\ t_{n_3 1} & \cdots & t_{n_3 a} & \cdots & t_{n_3 A} \\ \vdots & & \vdots & & \vdots \\ t_{N1} & \cdots & t_{Na} & \cdots & t_{NA} \end{bmatrix}$$

a = Walk

Formulation: Cost function

$$\min_{Z,T,w,b,v,c} \boxed{\frac{1}{N}\|Z-\phi(X)w-b\|_F^2+\lambda_1\, Tr(w^T\ w)}$$

$$+\boxed{\frac{1}{N}\|T-\psi(X)v-c\|_F^2+\lambda_2\, Tr(v^T\ v)}$$

Weak supervision from scripts:

Person p appears in clip N :

$$\sum_{n\in\mathcal{N}_i} z_{np} \geq 1$$

Action a appears in clip N :

$$\sum_{n\in\mathcal{N}_i} t_{na} \geq 1$$

Person p **and** Action a appear in clip N :

$$\sum_{n\in\mathcal{N}_i} z_{np}\ t_{na} \geq 1$$

Scaling to many movies: Faces

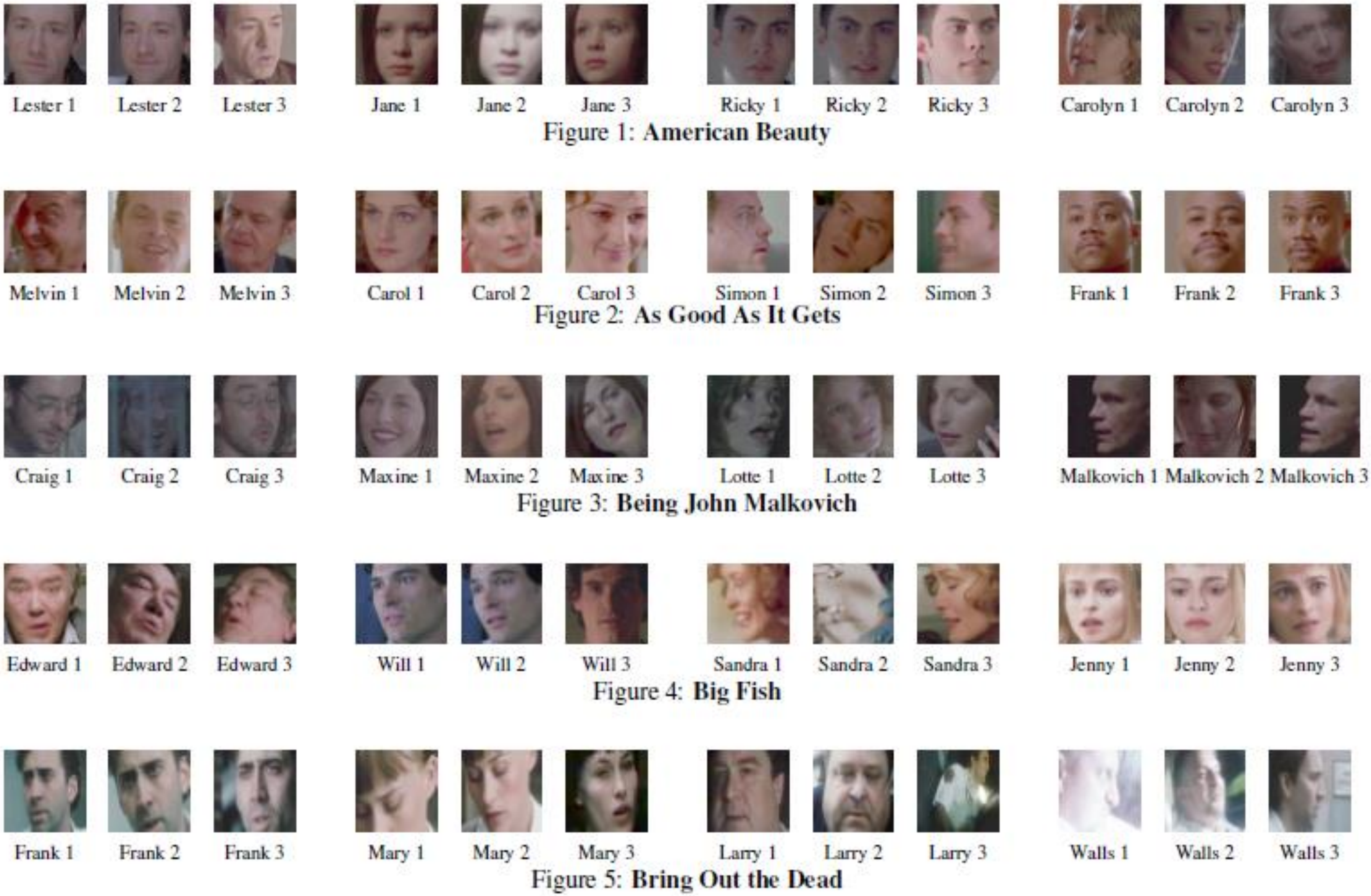


Figure 1: **American Beauty**

Figure 2: **As Good As It Gets**

Figure 3: **Being John Malkovich**

Figure 4: **Big Fish**

Figure 5: **Bring Out the Dead**

Scaling to many movies: Faces

Figure 9: **Charade**

Figure 10: **Chasing Amy**

Figure 11: **Clerks**

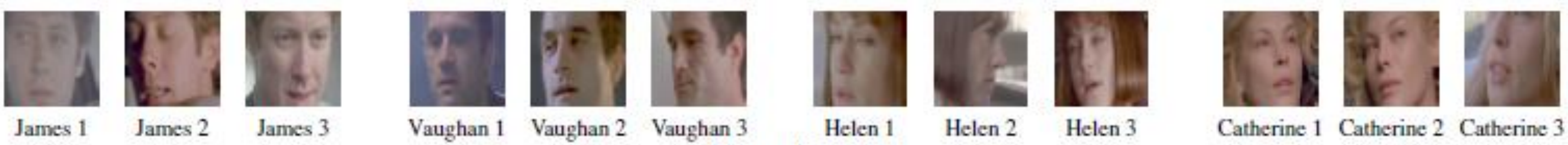


Figure 12: **Crash**

Figure 13: **Dead Poets Society**

Scaling to many movies: Faces

Figure 17: **Fargo**

Figure 18: **Fear and Loathing in Las Vegas**

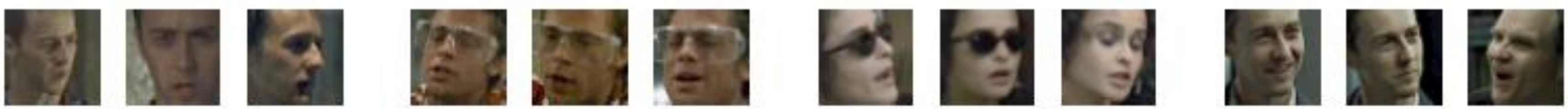


Figure 19: **Fight Club**

Figure 20: **Five Easy Pieces**

Figure 21: **Forrest Gump**


ROCKWELL
SPARTANS


P:REGGIE
A:AnswerPhone
Charade

How to define actions?

- Is action vocabulary well-defined?

Examples of “Open” action:






- What granularity of action vocabulary shall we consider?


Front Left

Source: http://www.youtube.com/watch?v=eYdUZdan5i8

Current solution: learn *person-throws-cat-into-trash-bin* classifier

What are action classes?

open

What is the right action granularity?

person-throws-cat-into-trash-bin

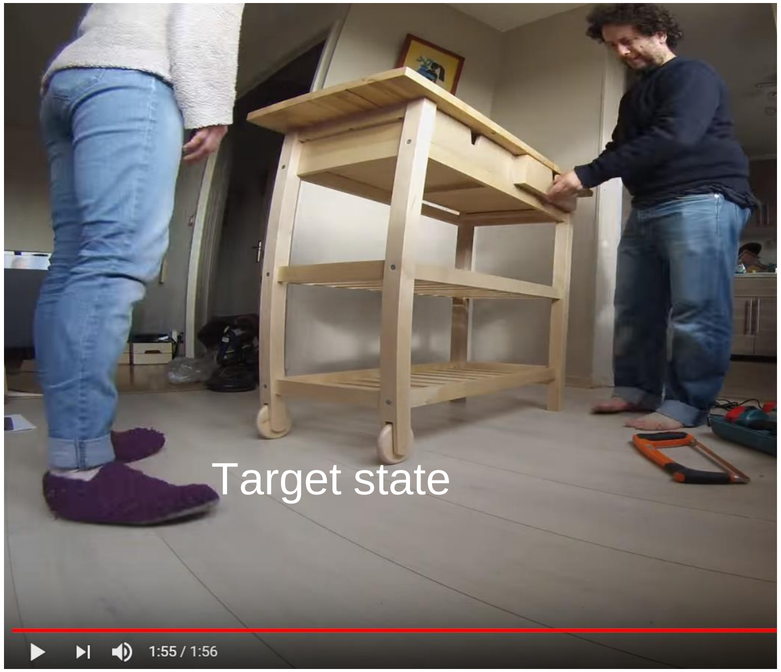





Define actions by *goals*

Instructional videos

- Narrated videos: people describe what they do
- Large variety of actions, objects, scenes and tasks
- Goal-driven sequences of actions



Learning from narrated instruction videos

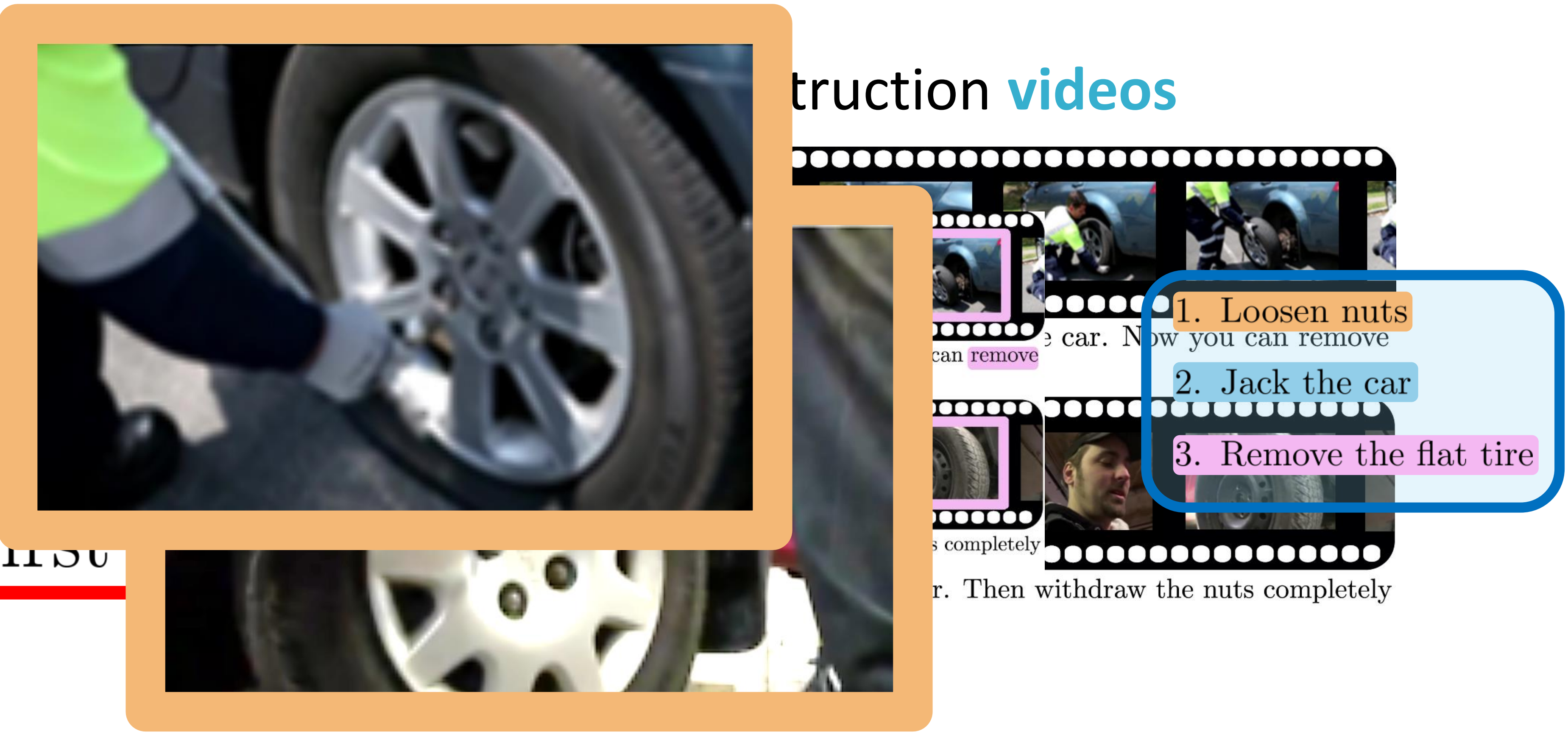
J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev and S. Lacoste-Julien

CVPR 2016

What are instructional videos?

[Alyarac et al., CVPR 2016]

How do we formalize the problem?



Outputs:
- sequence of **main steps**
- **visual** and **linguistic** representations of the steps
- temporal **localization** of each step

Assumptions and overview of the approach

Assumptions:

Assumption 1: Each task is composed of an **ordered** sequence of steps.

Assumption 2: People do **what** they say roughly **when** they say it

Approach:
two linked **clustering** stages

1) **Text clustering** using multiple sequence alignment

2) **Video clustering** under text constraints

Video clustering with text constraints

Text-based clustering

Video 1	Video 2	Video 3	Video 4
∅	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
∅	∅	unscrew nut	∅
remove wheel	remove tire	withdraw tire	∅
∅	lower jack	∅	lower car

Discovered list of steps

1) Loosen nut
2) Jack car
3) Remove wheel

Video clustering

$$h(Z) = \min_{W \in \mathbb{R}^{K \times d}} \underbrace{\frac{1}{2T} \|Z - XW\|_F^2}_{\text{Discriminative loss on data}} + \underbrace{\frac{\lambda}{2} \|W\|_F^2}_{\text{Regularizer}} .$$

OUTPUT (Discovered temporal localization)

Representation of video chunks (IDTF,CNN) [Txd] matrix

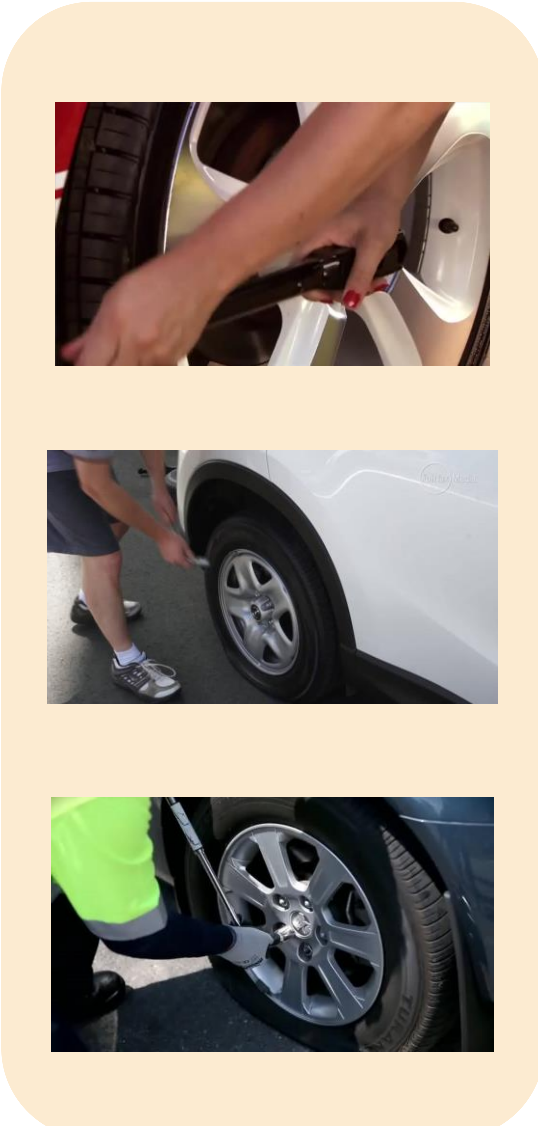
Linear action classifier [dxK] matrix

$$\min_{Z} \; h(Z) \quad \text{s.t.} \quad \underbrace{Z \in \mathcal{Z}}_{\text{ordered script}} \; , \quad \underbrace{AZ \geq R}_{\text{weak textual constraints}}$$

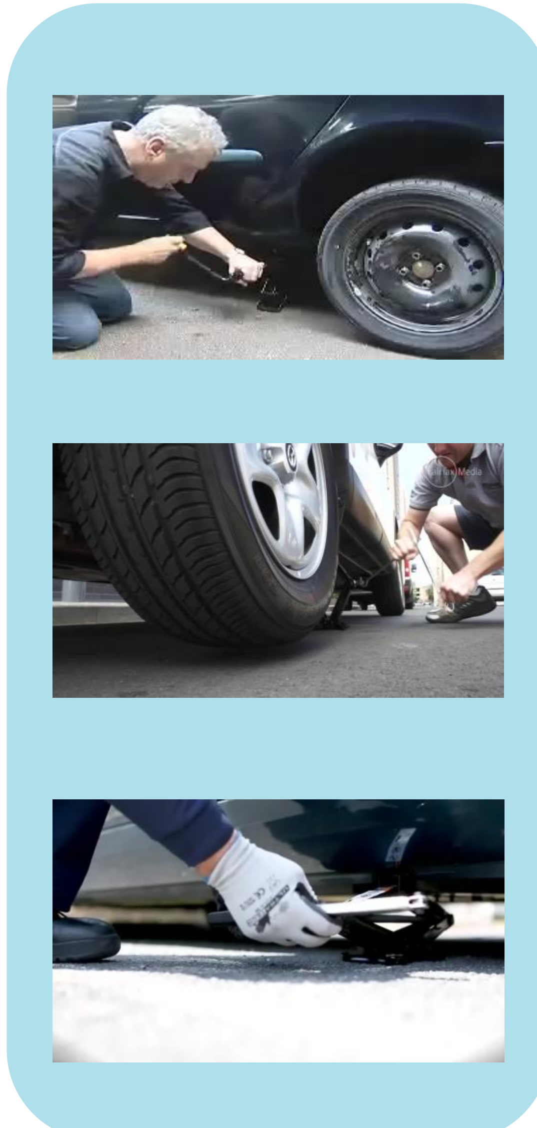
Subtitle alignment [SxT] matrix

Text Assignment [SxK] matrix

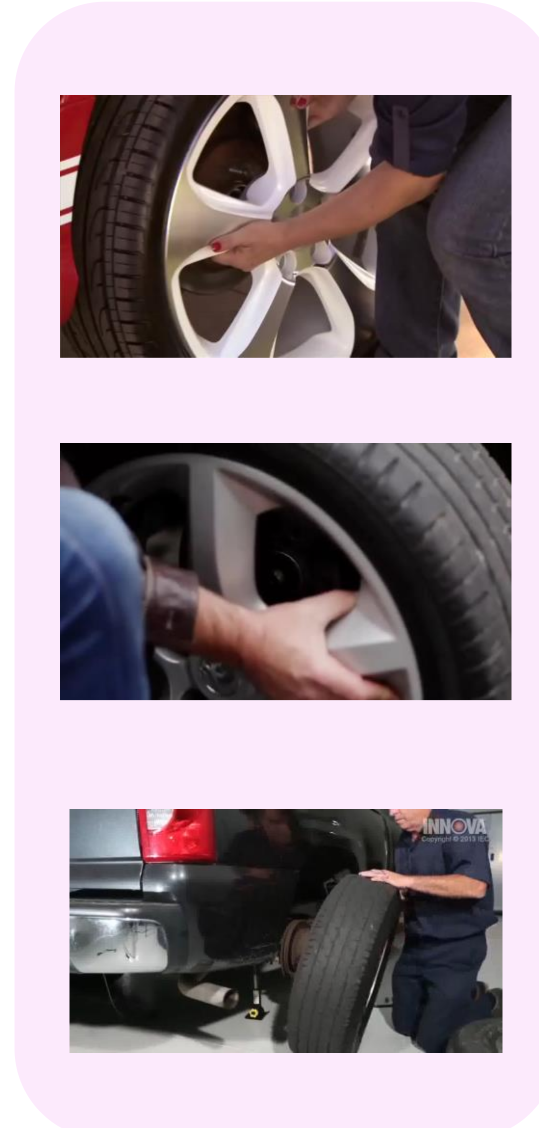
Qualitative results



``loosen nuts”



``jack car”

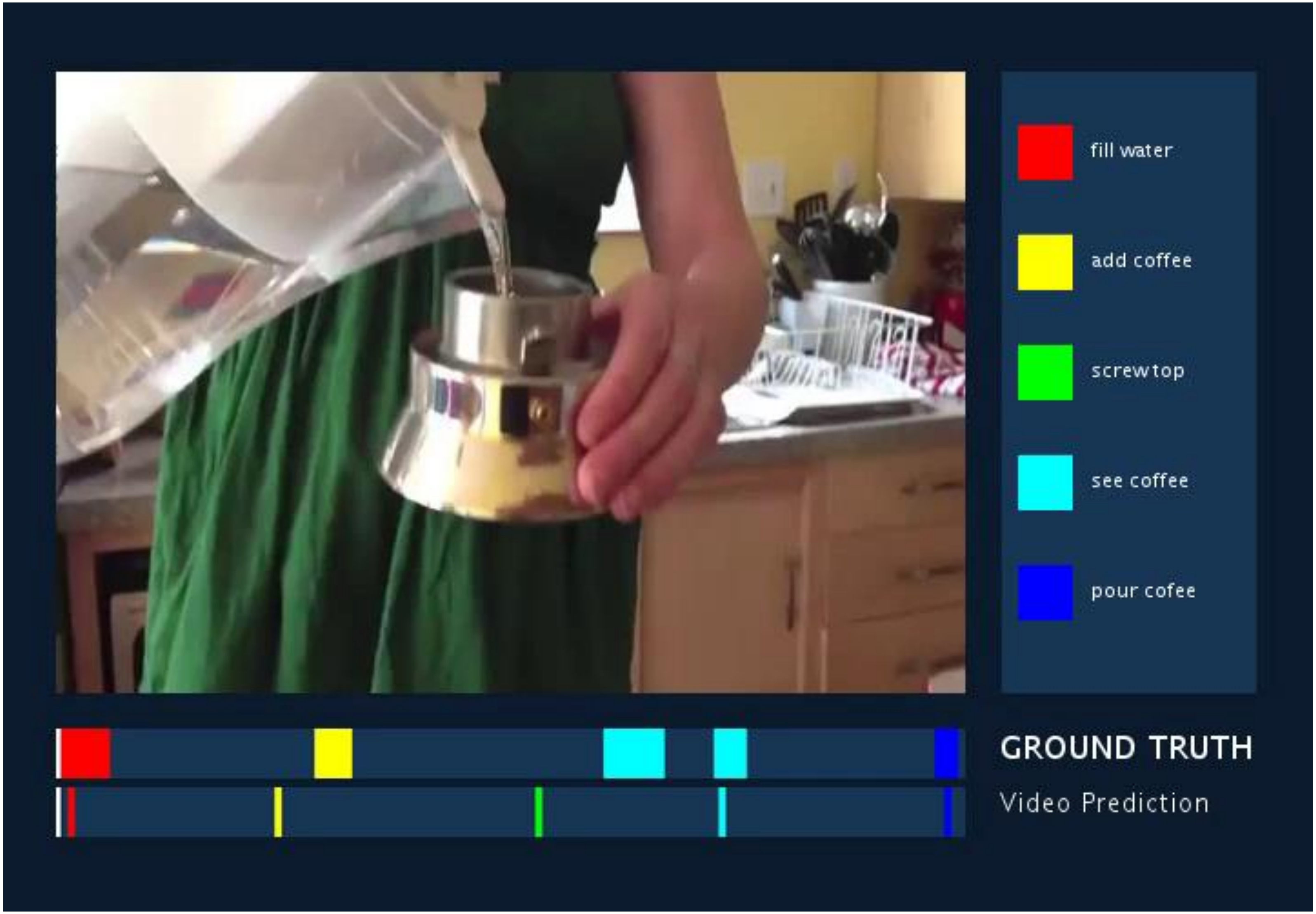


``remove wheel”

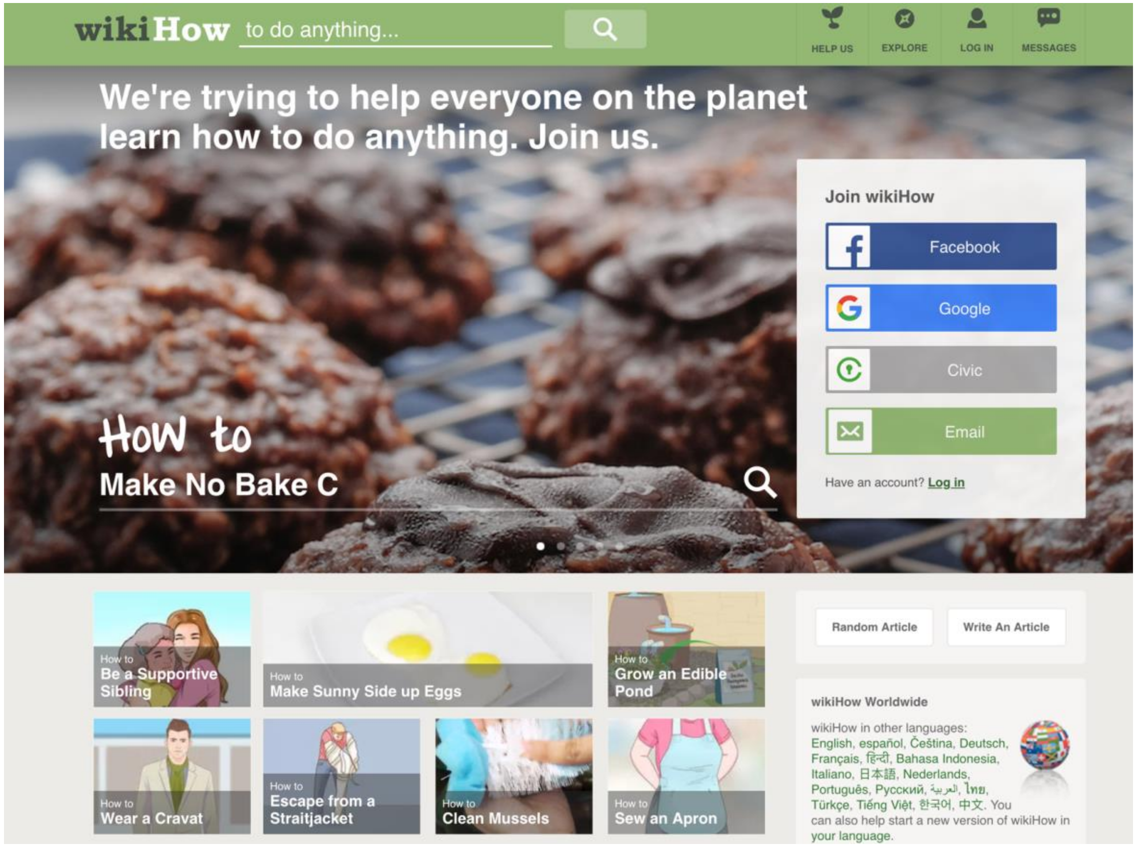
Changing a tire

K = 5

Qualitative results



Going WikiHow scale



Step 1: Scrap ~130K tasks from WikiHow

Examples of scrapped tasks

- ~~How to Be Healthy~~
- How to Cook Quinoa in a Rice Cooker
- How to Sew an Apron
- How to Break a Chain
- ~~How to April Fool your Girlfriend~~
- ....

Step 2: Filter out non-visual tasks

HowTo100M dataset

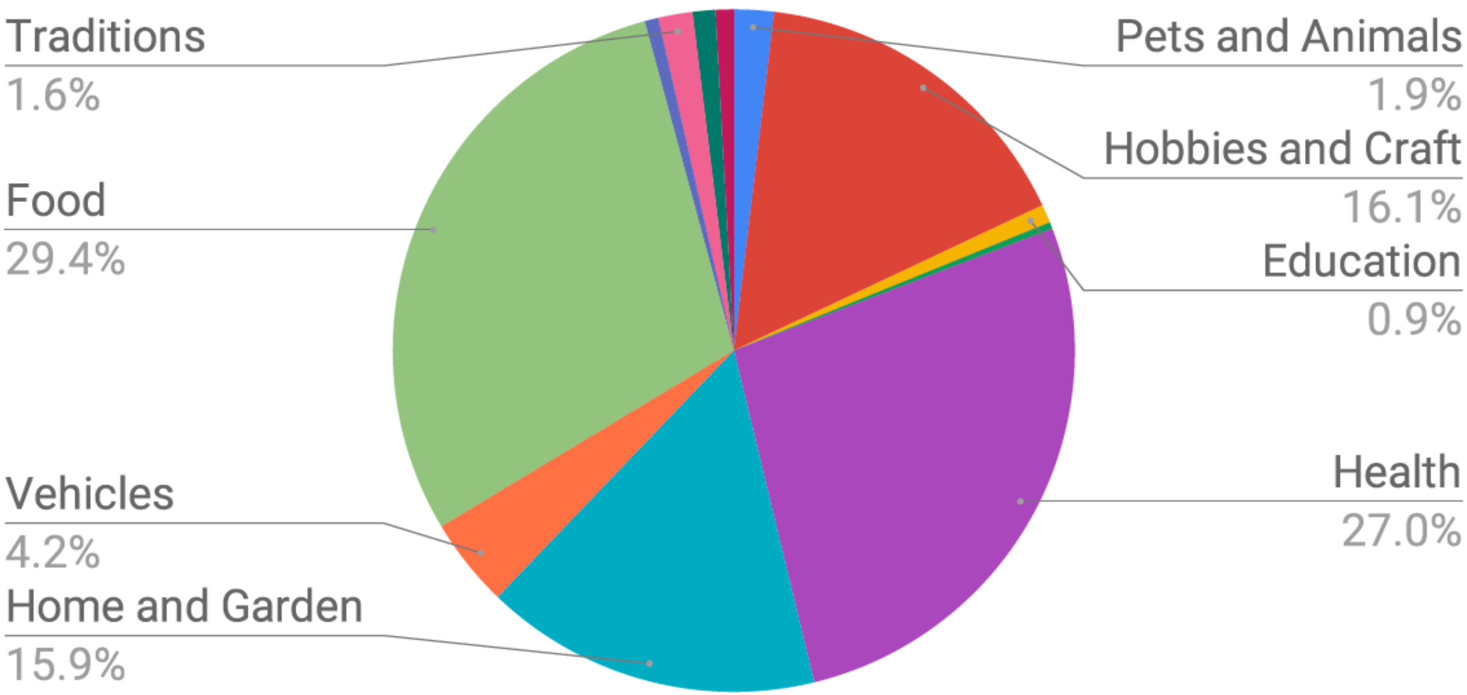
23 k — wikiHow visual tasks

1.2 M — YouTube videos

136 M — Short clips

ASR — Supervision
*ASR: Automated Speech Recognition

0 — Manual annotation



[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

HowTo100M dataset

23 k — wikiHow visual tasks

1.2 M — YouTube videos

136 M — Short clips

ASR — Supervision
*ASR: Automated Speech Recognition

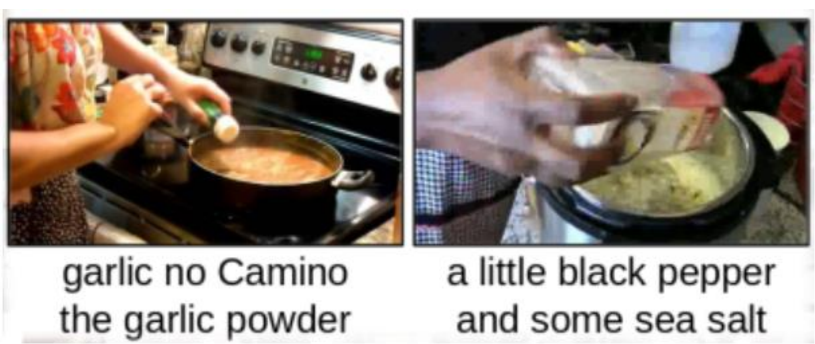
0 — Manual annotation

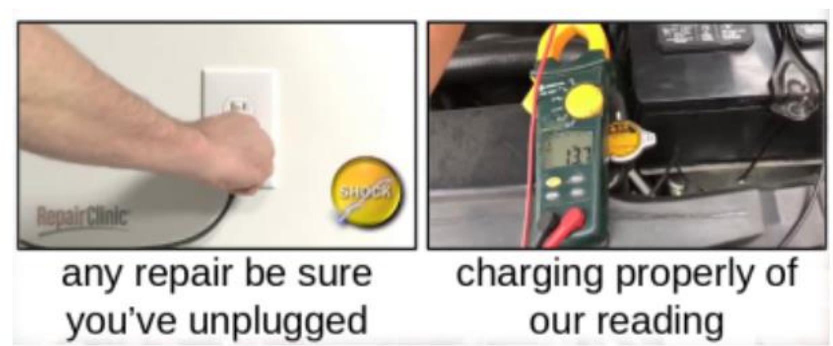
Dataset	Clips	Captions	Videos	Duration	Source	Year
Charades	10k	16k	10,000	82h	Home	2016
MSR-VTT	10k	200k	7,180	40h	Youtube	2016
YouCook2	14k	14k	2,000	176h	Youtube	2018
EPIC-KITCHENS	40k	40k	432	55h	Home	2018
DiDeMo	27k	41k	10,464	87h	Flickr	2017
M-VAD	49k	56k	92	84h	Movies	2015
MPII-MD	69k	68k	94	41h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

x1000 times larger!

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

HowTo100M dataset: Examples



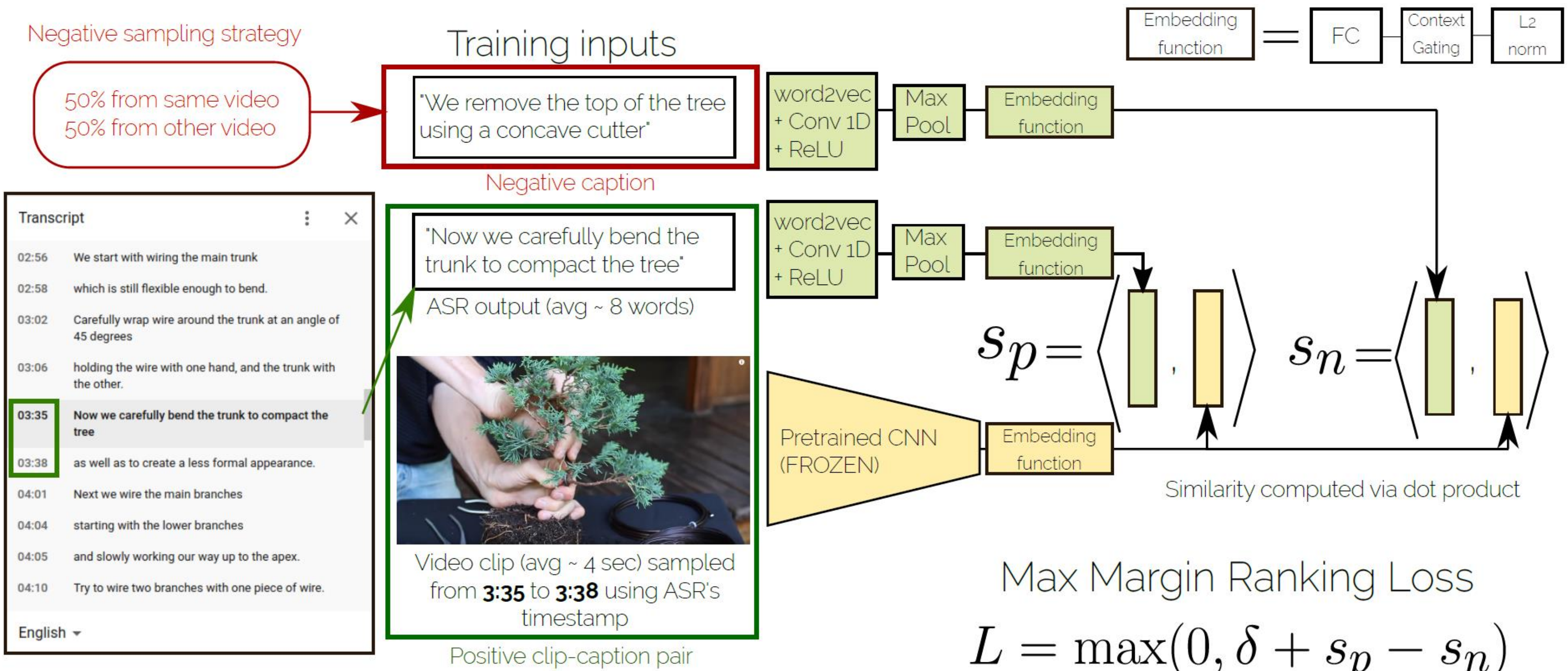


[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

Joint embedding model

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

Joint embedding model



[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

Online search demo: https://www.di.ens.fr/willow/research/howto100m/

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

Results: Instructional videos

YouCook2: Video retrieval

Method	Trainset	R@1	R@5	R@10	Median R
Random	None	0.03	0.15	0.3	1675
HGLMM FV CCA [21]	YouCook2	4.6	14.3	21.6	75
Ours	YouCook2	4.2	13.7	21.5	65
Ours	HowTo100M	6.1	17.3	24.8	46
Ours	PT: HowTo100M FT: YouCook2	**8.2**	**24.5**	**35.3**	**24**

Weakly-supervised training on HowTo100M outperforms fully-supervised training on YouCook2 and CrossTask datasets

Fine-tuning gives further improvements

CrossTask: Action localization

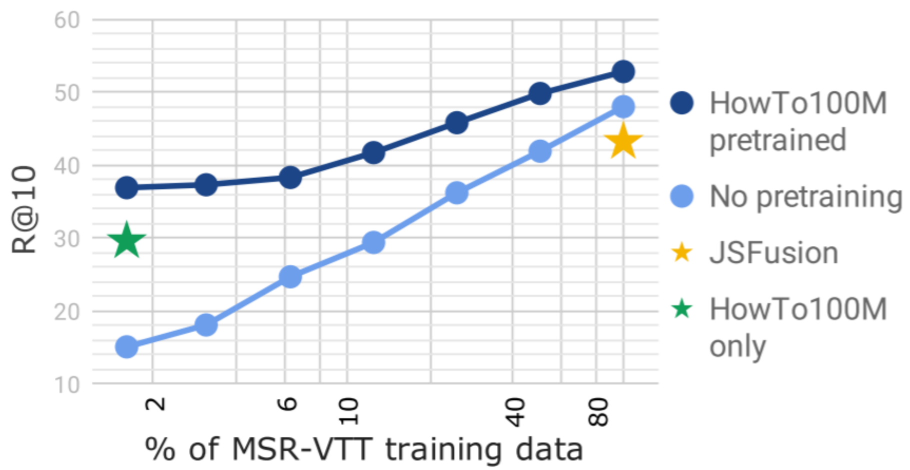
	Make Kimchi Rice	Pickle Cucumber	Make Banana Ice Cream	Grill Steak	Jack Up Car	Make Jello Shots	Change Tire	Make Lemonade	Add Oil to Car	Make Latte	Build Shelves	Make Taco Salad	Make French Toast	Make Irish Coffee	Make Strawberry Cake	Make Pancakes	Make Meringue	Make Fish Curry	Average
Fully-supervised upper-bound [68]	19.1	25.3	38.0	37.5	25.7	28.2	54.3	25.8	18.3	31.2	47.7	12.0	39.5	23.4	30.9	41.1	53.4	17.3	31.6
Alayrac *et al.* [2]	15.6	10.6	7.5	14.2	9.3	11.8	17.3	13.1	6.4	12.9	27.2	9.2	15.7	8.6	16.3	13.0	23.2	7.4	13.3
Zhukov *et al.* [68]	13.3	18.0	23.4	23.1	16.9	16.5	30.7	21.6	4.6	19.5	35.3	10.0	32.3	13.8	29.5	37.6	**43.0**	13.3	22.4
Ours trained on HowTo100M only	**33.5**	**27.1**	**36.6**	**37.9**	**24.1**	**35.6**	**32.7**	**35.1**	**30.7**	**28.5**	**43.2**	**19.8**	**34.7**	**33.6**	**40.4**	**41.6**	41.9	**27.4**	**33.6**

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

Results: YouTube videos

MSR-VTT: Video retrieval

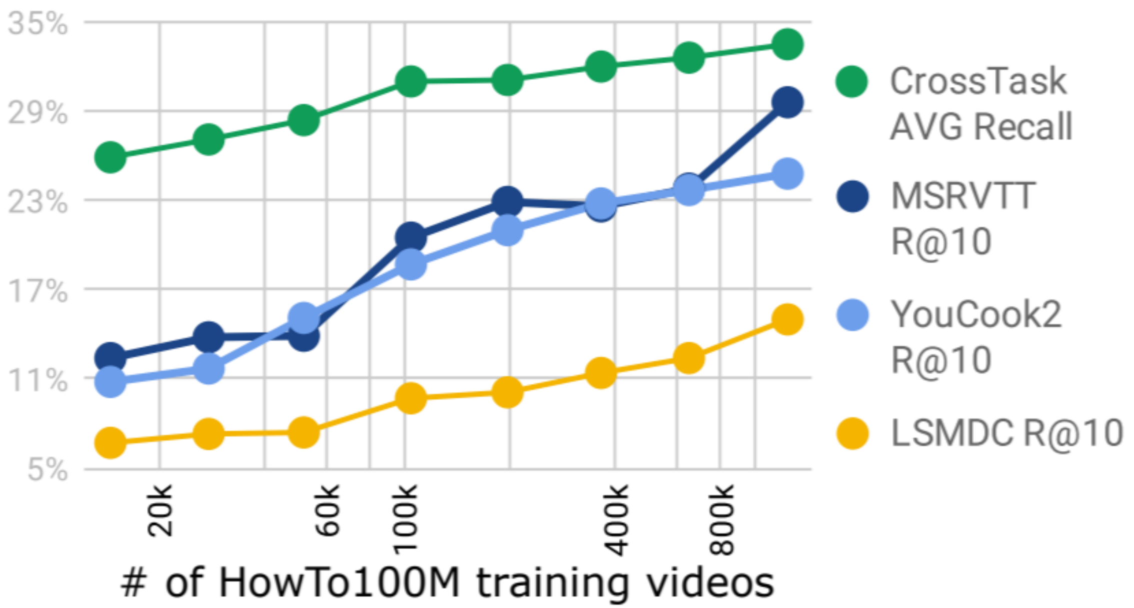
Method	Trainset	R@1	R@5	R@10	Median R
Random	None	0.1	0.5	1.0	500
C+LSTM+SA+FC7 [47]	MSR-VTT	4.2	12.9	19.9	55
VSE-LSTM [20]	MSR-VTT	3.8	12.7	17.1	66
SNUVL [58]	MSR-VTT	3.5	15.9	23.8	44
Kaufman *et al*. [18]	MSR-VTT	4.7	16.6	24.1	41
CT-SAN [59]	MSR-VTT	4.4	16.6	22.3	35
JSFusion [57]	MSR-VTT	10.2	31.2	43.2	13
Ours	MSR-VTT	12.1	35.0	48.0	12
Ours	HowTo100M	7.5	21.2	29.6	38
Ours	PT: HowTo100M FT: MSR-VTT	**14.9**	**40.2**	**52.8**	**9**



Pre-training on HowTo100M + finetuning outperforms fully-supervised training on MSR-VTT

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

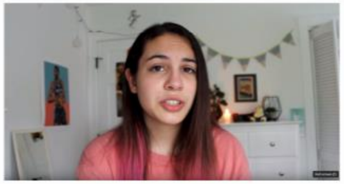
Results: Impact of scale



[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

Open challenges

- HowTo100M contains ~50% label noise due to video-text misalignment, non-visual explanations, etc.



... want to be that extra right when you finish a question ...

... by our electronic devices and in the same cases in your plants ...

... on my old moto Guzzi or had before I sold it ...

- Our method relies on pre-trained video features, no end-to-end learning of visual representations despite massive (but noisy) data.

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



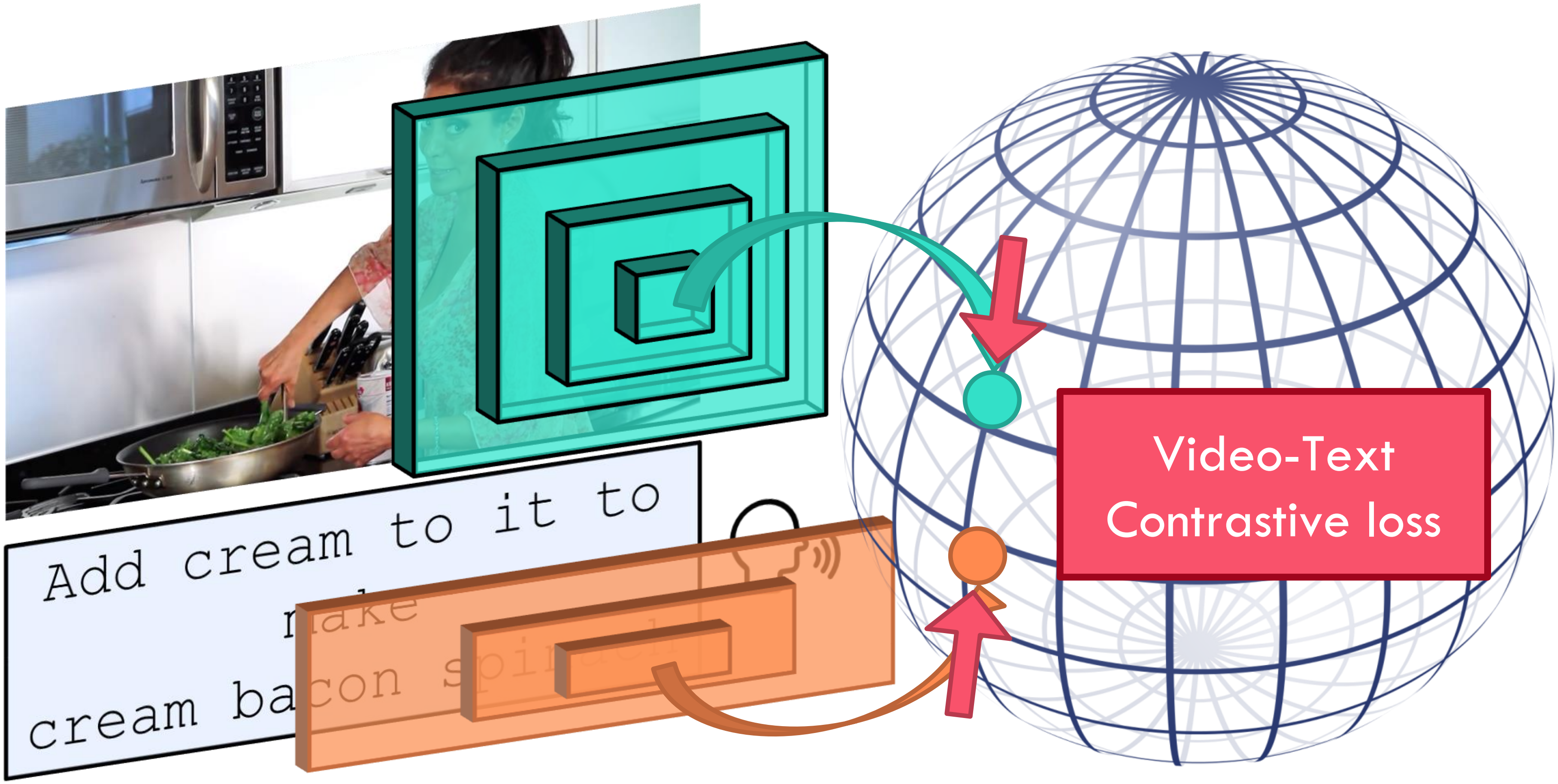



A. Miech*, J-B. Alayrac*, L. Smaira, I. Laptev, J. Sivic, A. Zisserman

*equal contributions

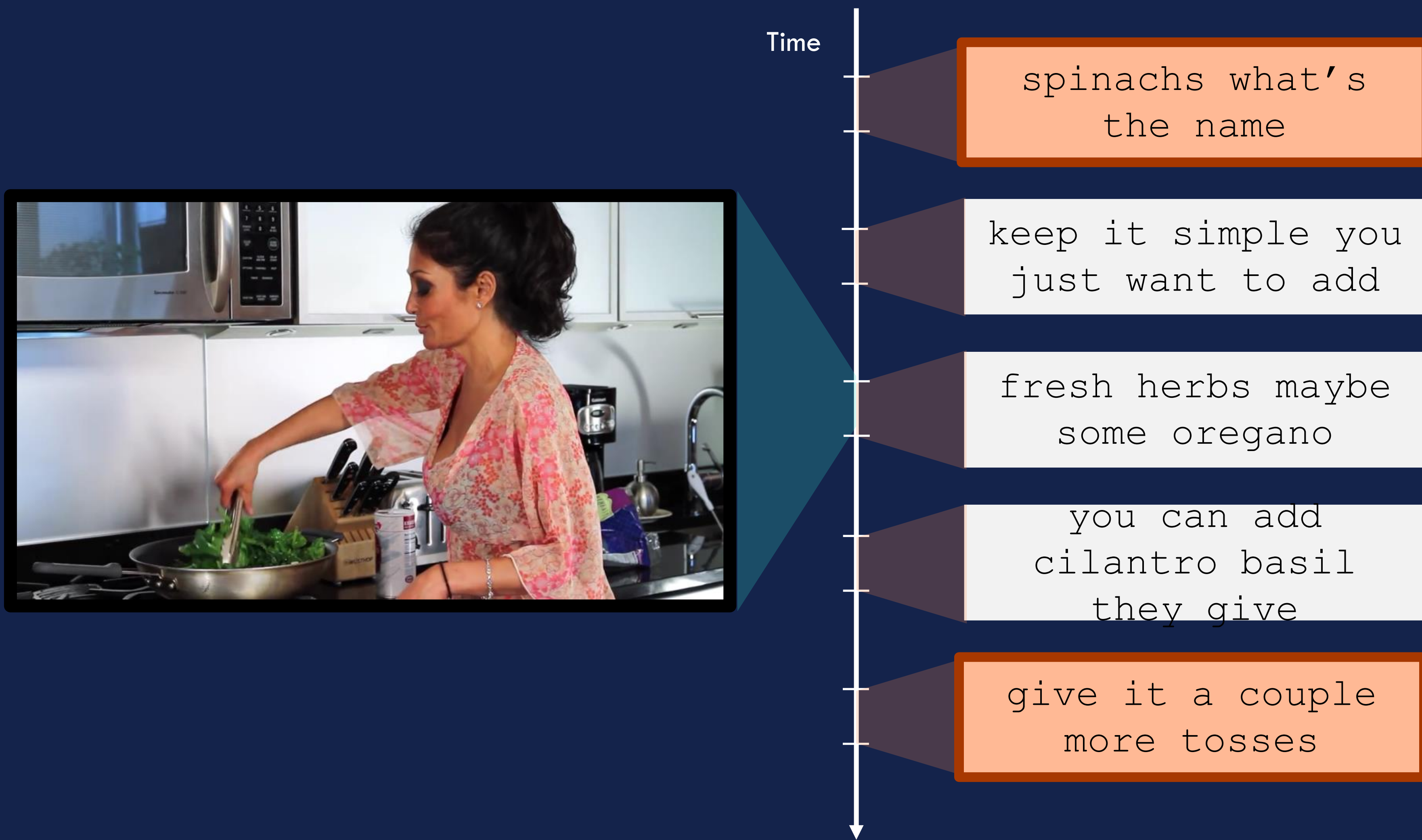
CVPR 2020

Training task

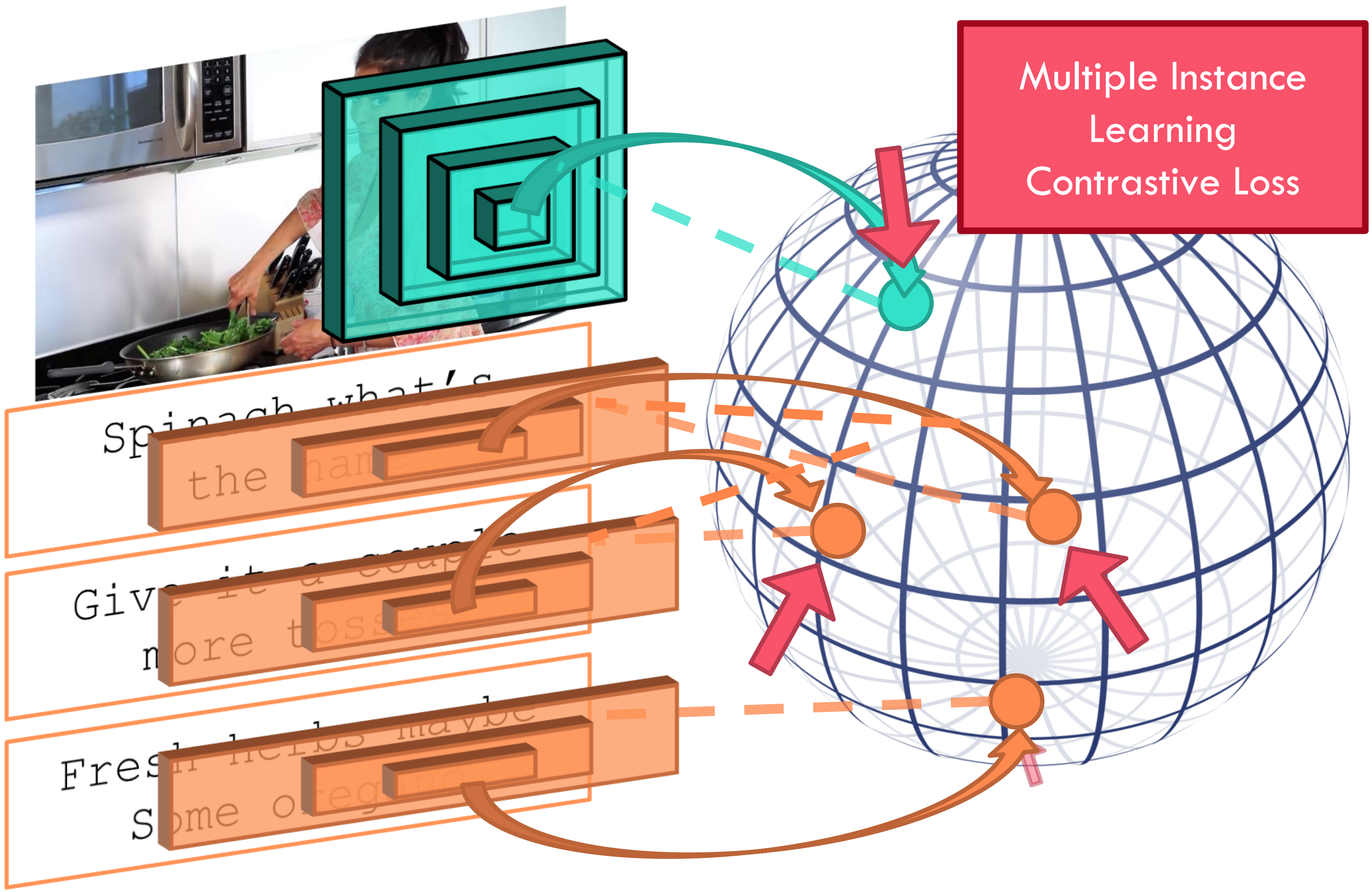


[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]



[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]



[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

Our formulation: **MIL-NCE**

Video input

Narration input

$$e^{f(x)^\top g(y)}$$

f, g

Video network

Text network

Bag of positive candidate pairs

Negative video-narration pairs

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

Our formulation: **MIL-NCE**

Spinach what's the name

Give it a couple more tosses

Fresh herbs maybe Some oregano

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

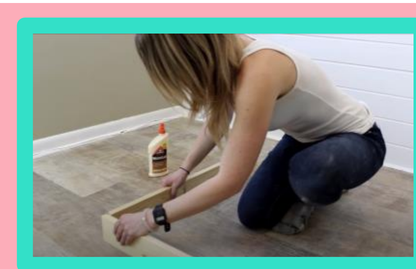
Bag of positive candidate pairs

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

Our formulation: **MIL-NCE**

Let's glue the
piece of woods

Keep it simple
you
Just want to add



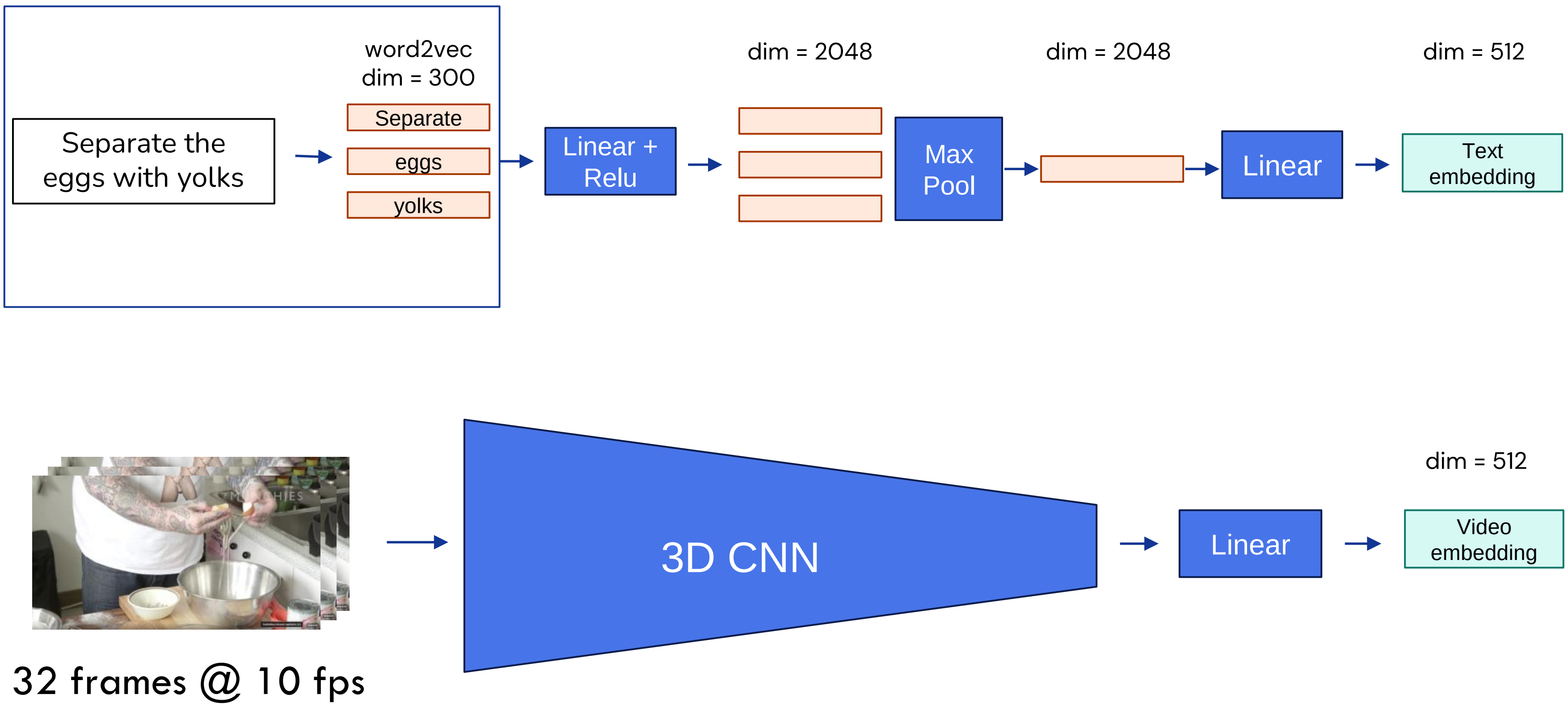
Fresh herbs
maybe
Some oregano

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

Negative
video-narration pairs

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

Video-Text model architecture



[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

The downstream tasks

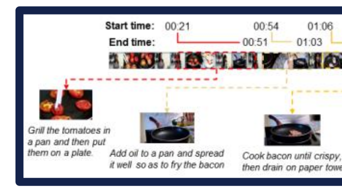
Action recogniti on

HMDB-51

UCF-101

Text-to-Video retrieval

MSR-VTT



YouCook2

Action Localization

COIN
A large-scale dataset for COmprehensive INstructional video analysis

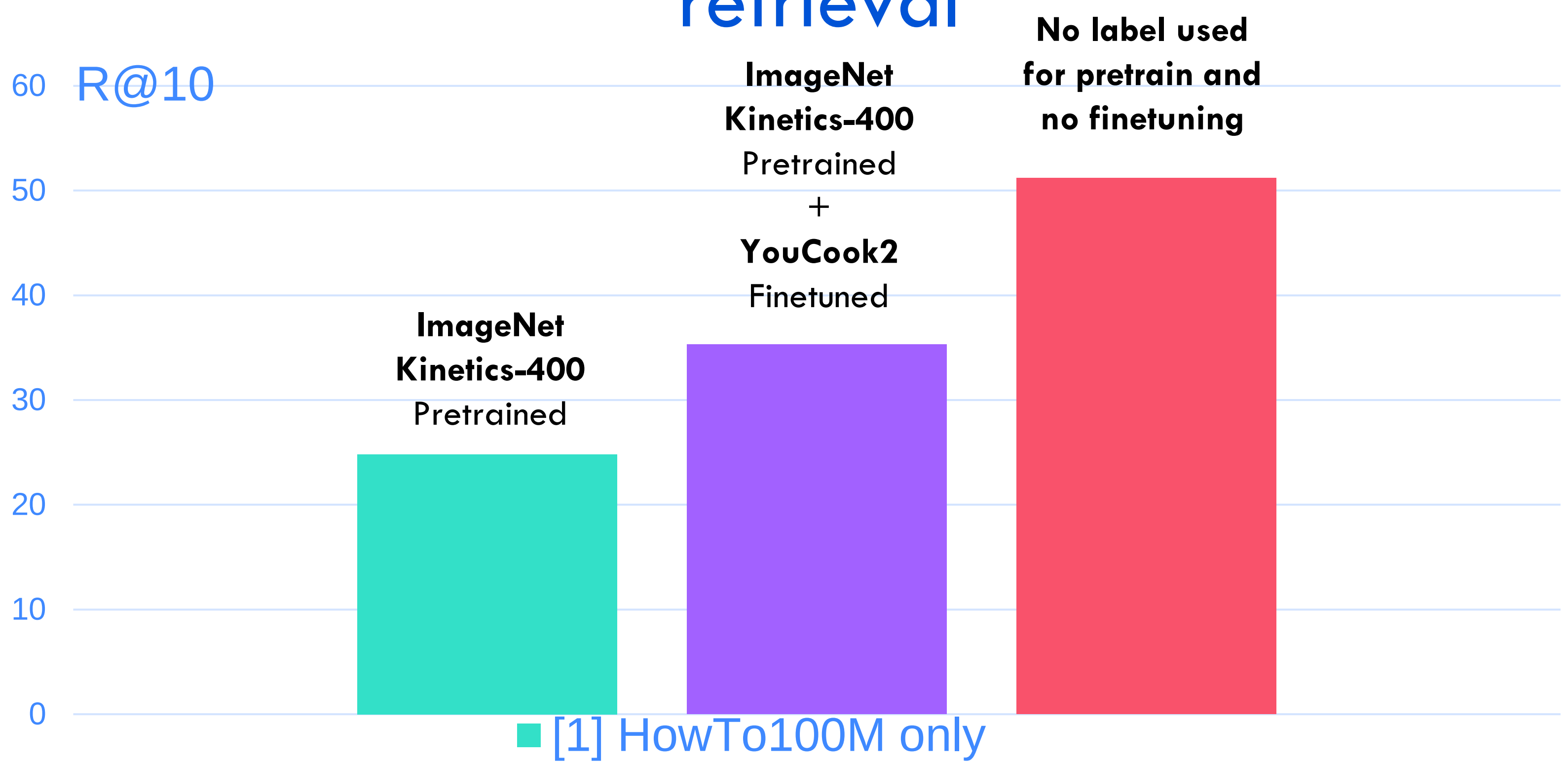
COIN

YouTube 8M Segments

CrossTask

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

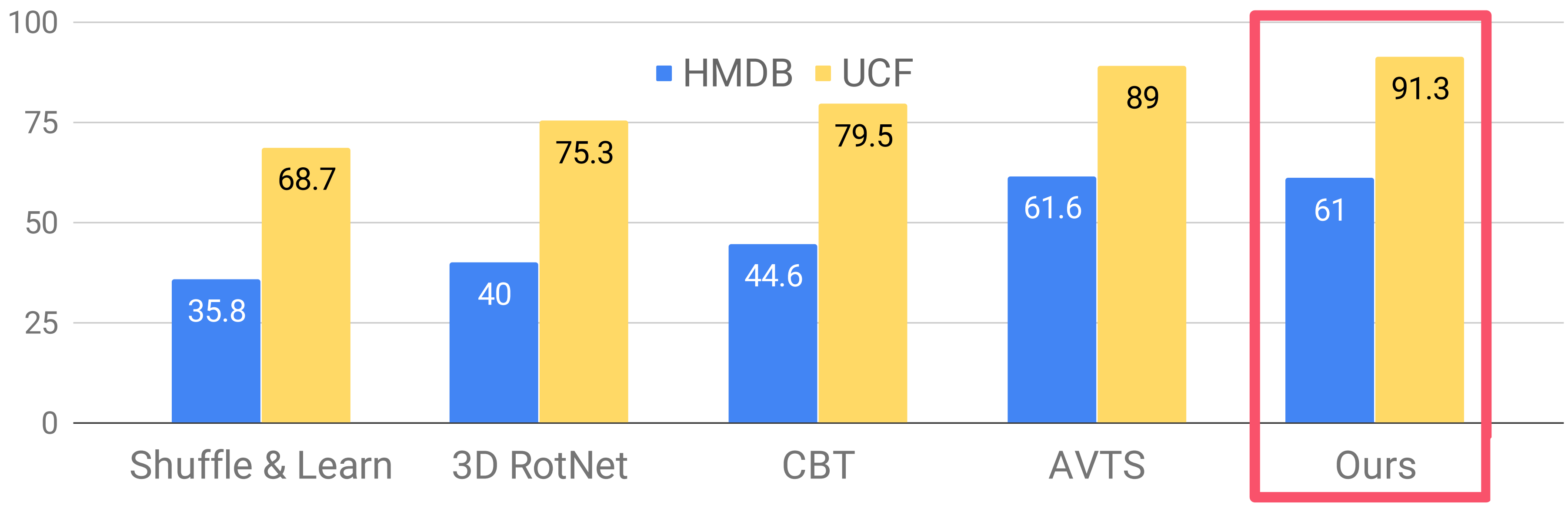
YouCook2 Zero-Shot Text-to-Video retrieval



[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.

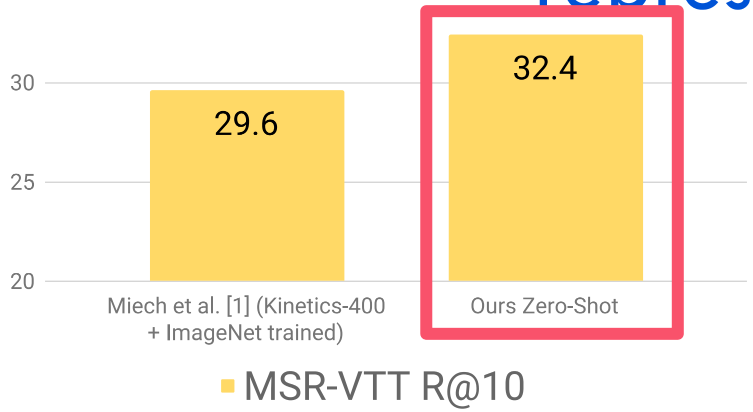
[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

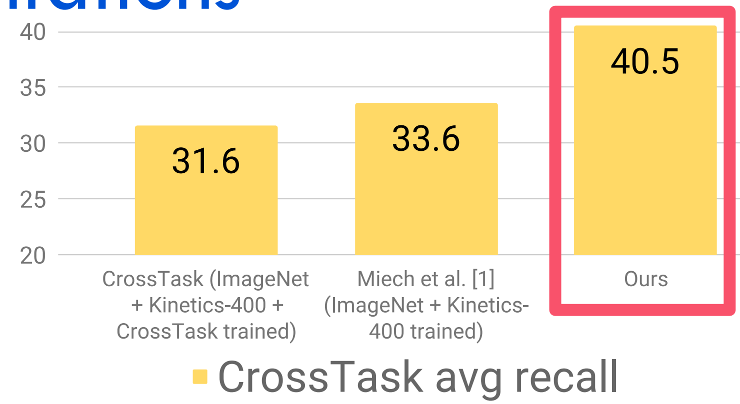
Action recognition: comparison to self-supervised video representations

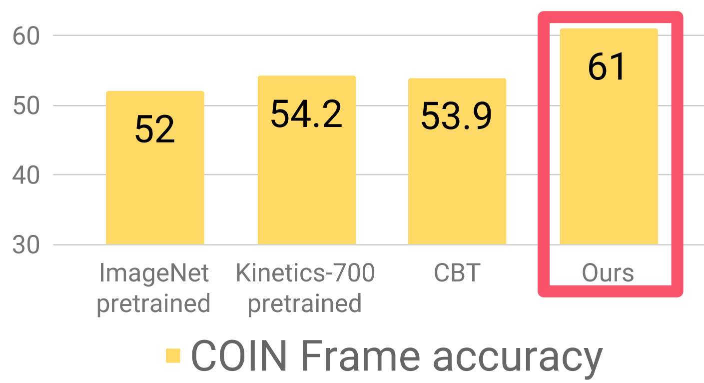


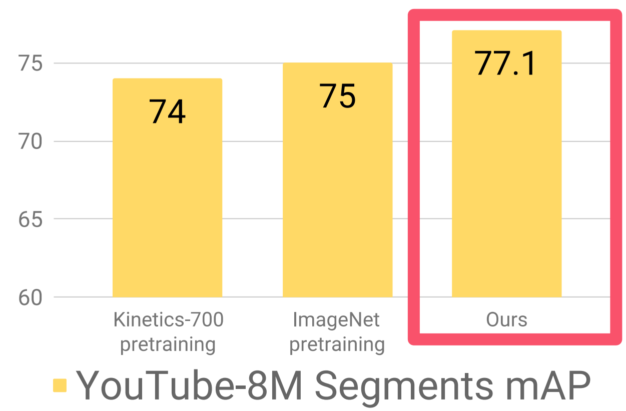
[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

Comparison to fully-supervised representations









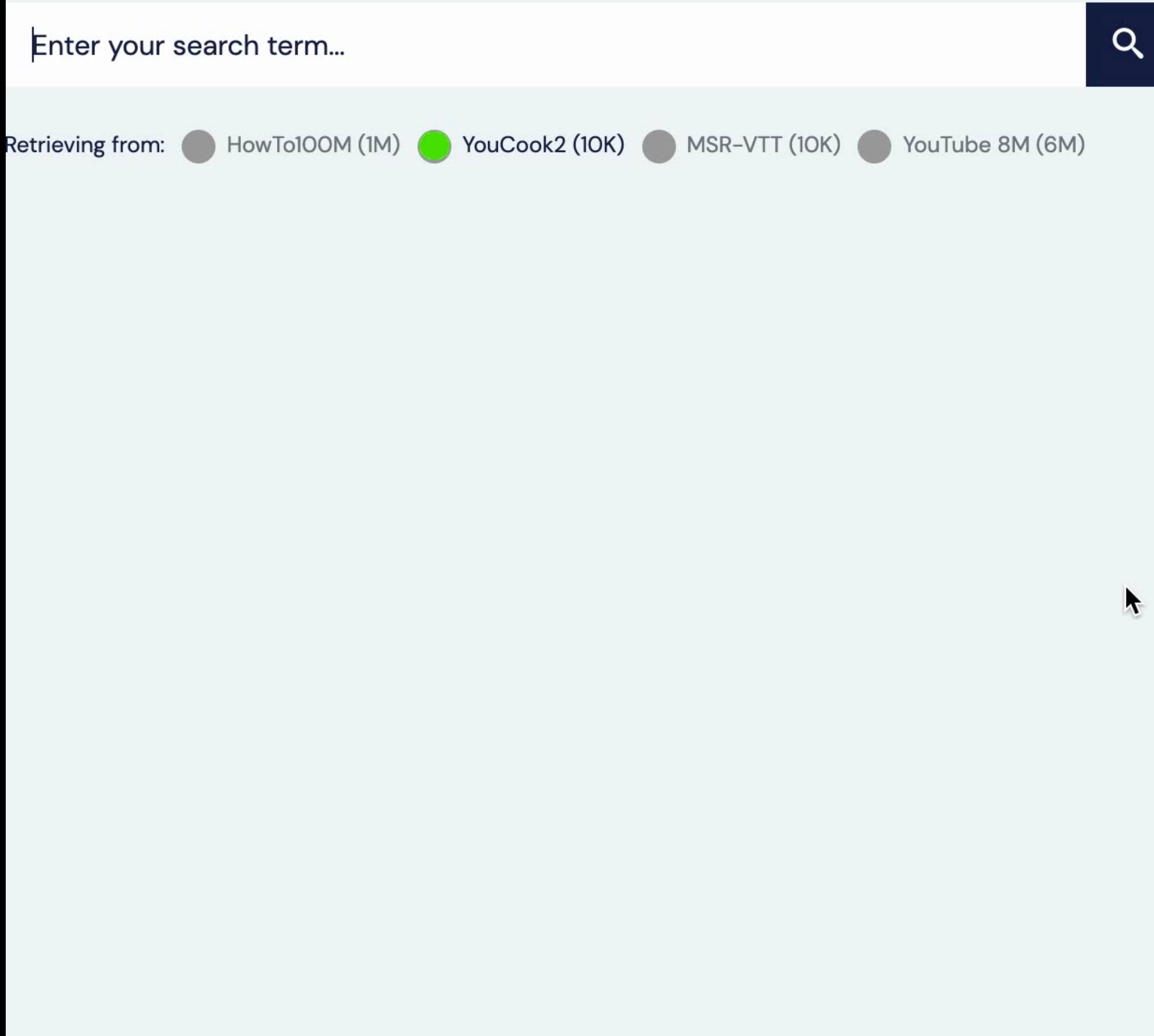
[1] *A.* Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic,
HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, in ICCV, 2019.

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

Pretrained Text-Video models and code publicly available

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

Video search by text



Recent work on learning from images and text

OpenAI CLIP:

Radford et al., Learning transferable visual models from natural language supervision. arXiv:2103.00020. 2021 Feb 26.

Microsoft:

Yuan et al., Florence: A New Foundation Model for Computer Vision. arXiv preprint arXiv:2111.11432. 2021 Nov 22.

Google:

Jia et al., Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918. 2021 Feb 11.

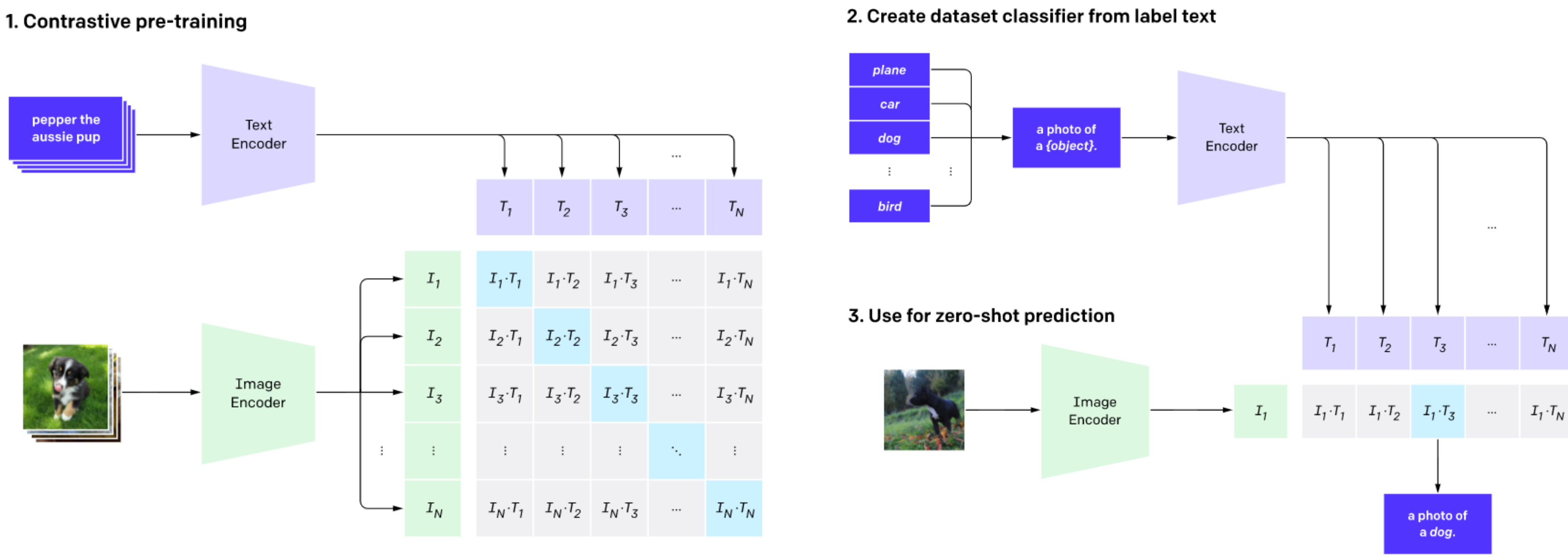
Pham et al., Combined Scaling for Zero-shot Transfer Learning. arXiv preprint arXiv:2111.10050. 2021 Nov 19.

Recent work on learning from images and text

OpenAI CLIP:

Radford et al., Learning transferable visual models from natural language supervision. arXiv:2103.00020. 2021 Feb 26.

Training on 400M pairs of images and text



Recent work on learning from images and text

Pham et al., Combined Scaling for Zero-shot Transfer Learning. arXiv preprint arXiv:2111.10050. 2021 Nov 19.

	ALIGN [37]	**CLIP** [64]	**BASIC** (ours)
ImageNet	76.4	76.2	**85.7** (+9.3)
ImageNet-A	75.8	77.2	**85.6** (+8.4)
ImageNet-R	92.2	88.9	**95.7** (+3.5)
ImageNet-V2	70.1	70.1	**80.6** (+10.5)
ImageNet-Sketch	64.8	60.2	**76.1** (+11.3)
ObjectNet	72.2	72.3	**78.9** (+6.6)
Average	74.5	74.2	**83.7** (+9.2)

Table 1: Highlights of our key results. Shown are the top-1 accuracy of our method, BASIC, and other state-of-the-art zero-shot transfer methods – CLIP and ALIGN – on ImageNet and other robustness test sets. None of these models has seen any ImageNet training example. On average, BASIC surpasses these methods by the significant 9.2 percentage points.

Adapting Large Language Models

- Paired Vision-Language data on the Internet is
  - (a) Noisy and
  - (b) Relatively scarce compared to Language-only data
- Large Language Models (LLMs) already encode much of the common-sense knowledge that could be useful for vision tasks.

(Some) recent work adopting LLMs for vision tasks:

- Brown et al., Language models are few-shot learners. *In Proc NeurIPS 2020*.
- Alayrac et al., Flamingo: a visual language model for few-shot learning. *In Proc NeurIPS 2022*.
- Li et al., Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *In Proc ICML 2023.*
- Liu et al., Visual Instruction Tuning. In Proc NeurIPS 2023
- …

Flamingo: a Visual Language Model for Few-Shot Learning

Alayrac et al., NeurIPS 2022

Architecture

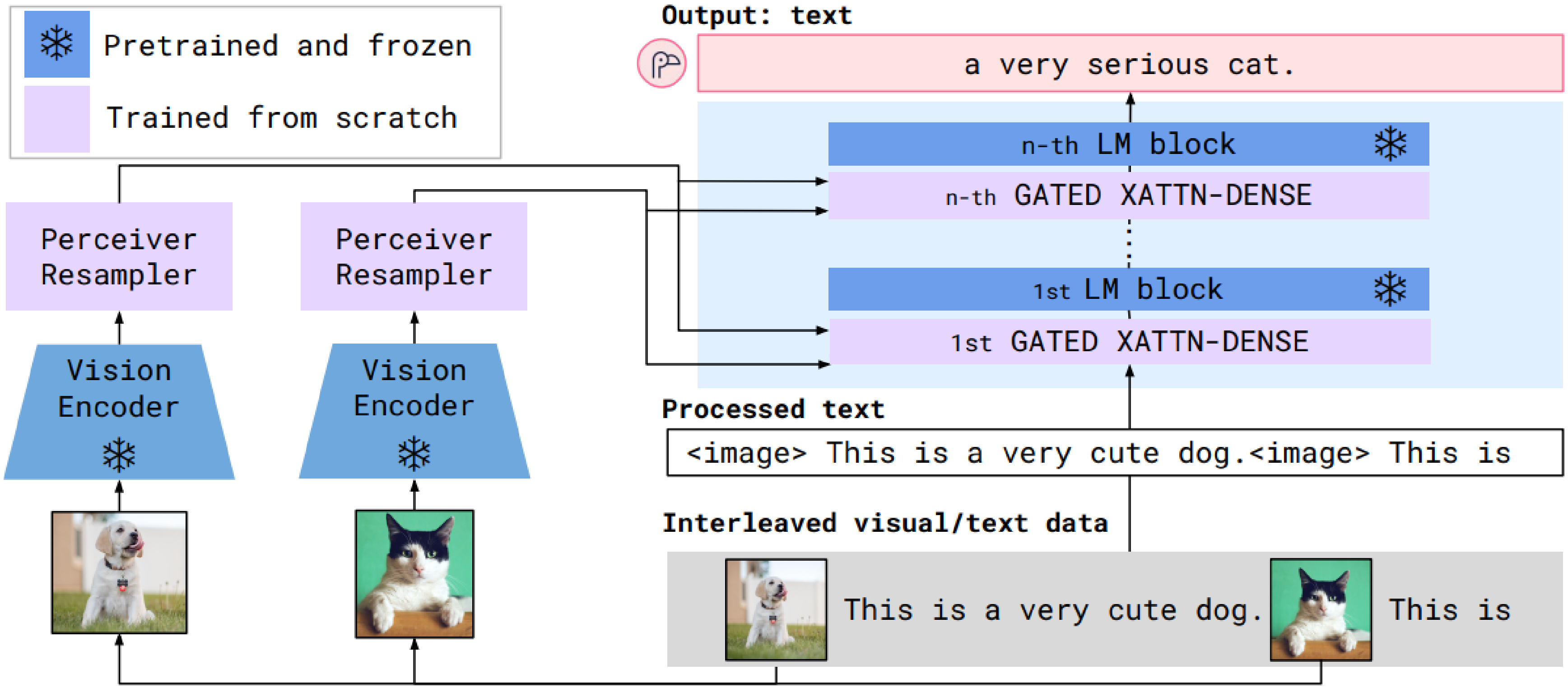


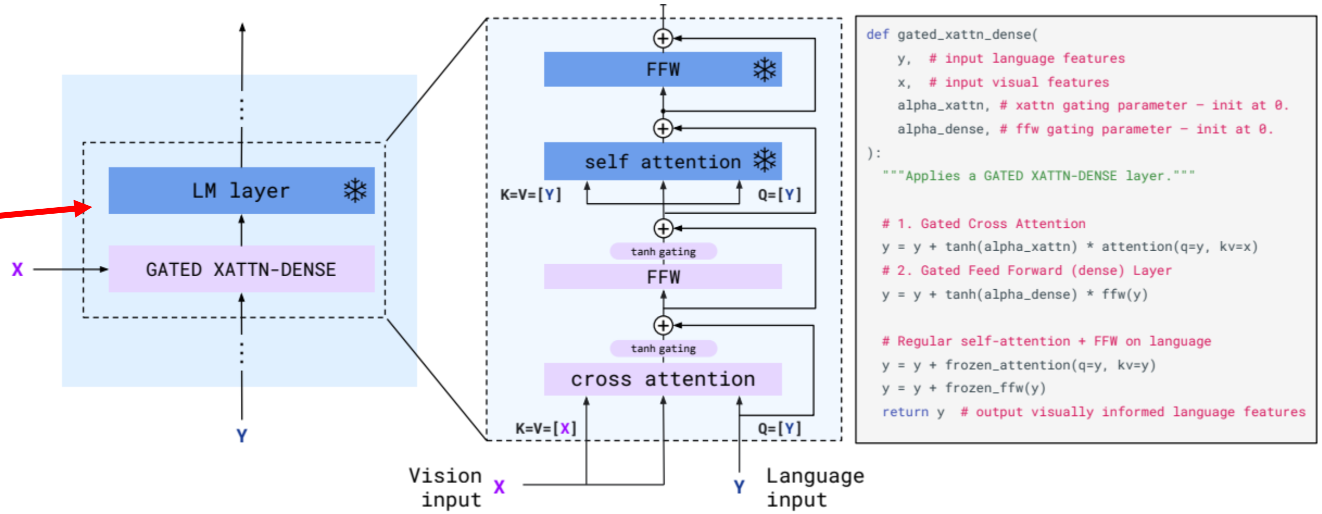
Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

🦩 Flamingo: a Visual Language Model for Few-Shot Learning

Alayrac et al., NeurIPS 2022

Training

Pre-trained
70B-parameter
LLM Chinchilla



Vision-Language training data:

- MultiModal MassiveWeb (M3W) dataset obtained from 43M webpages
- ALIGN dataset with 1.8B images paired with alt-text.
- VTP (Video & Text Pairs) with 27M short videos paired with sentence descriptions

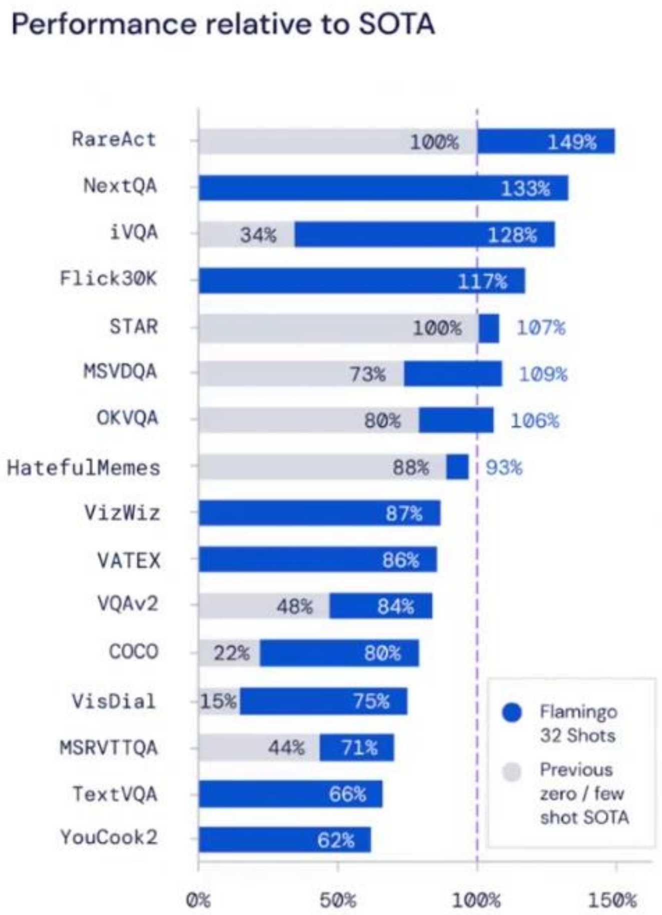
Training objective:

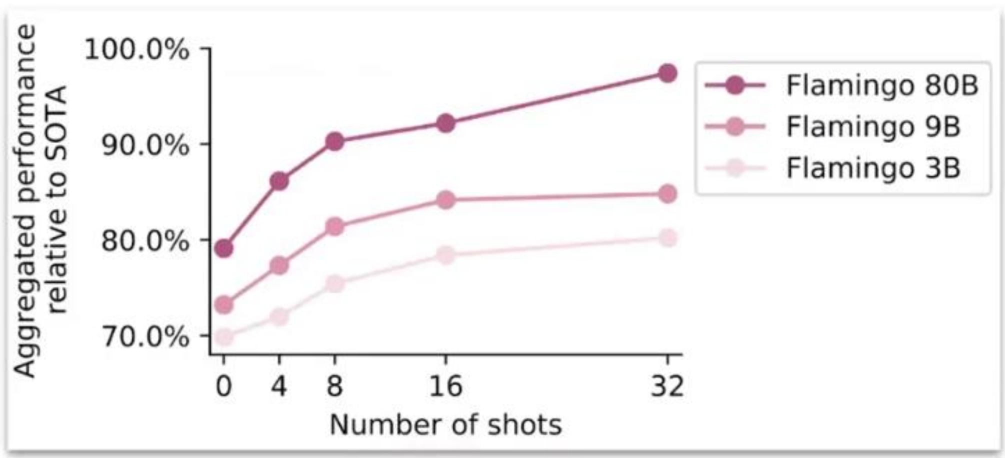
- Text prediction givenvisual input

🦩 Flamingo: a Visual Language Model for Few-Shot Learning

Alayrac et al., NeurIPS 2022

Results

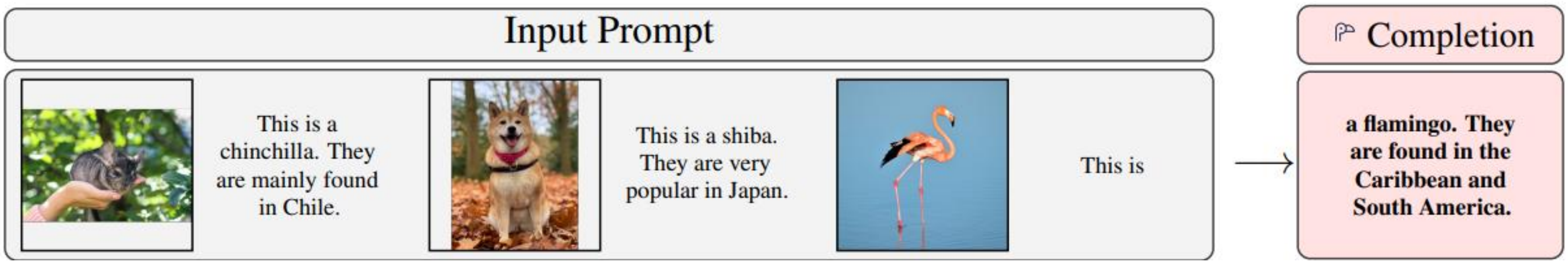




Flamingo: a Visual Language Model for Few-Shot Learning

Alayrac et al., NeurIPS 2022

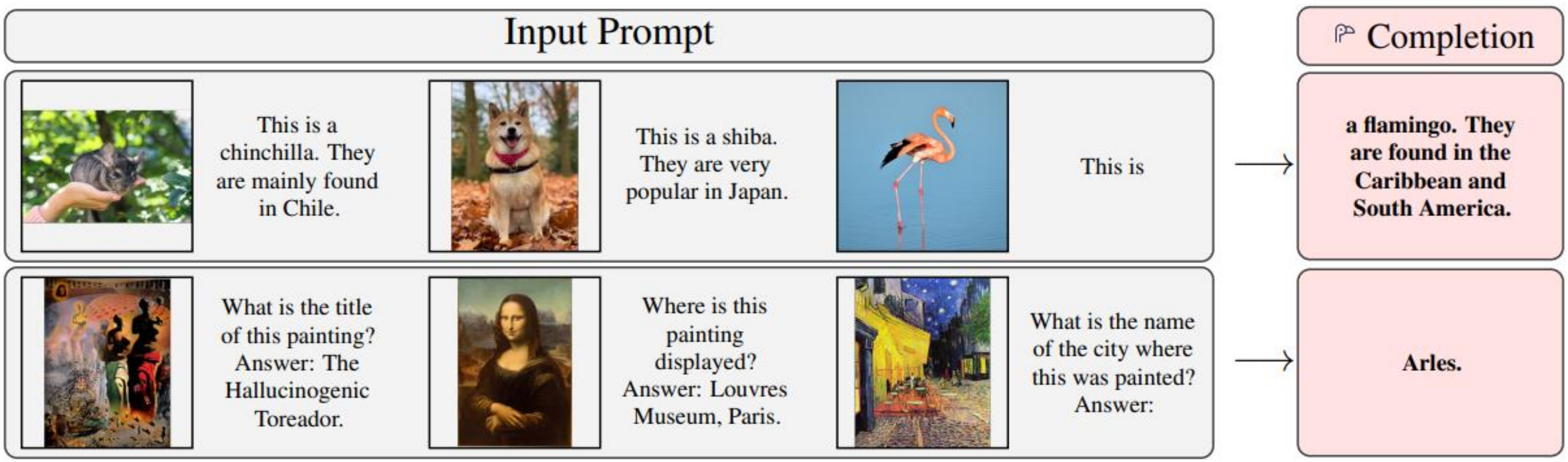
Results



Flamingo: a Visual Language Model for Few-Shot Learning

Alayrac et al., NeurIPS 2022

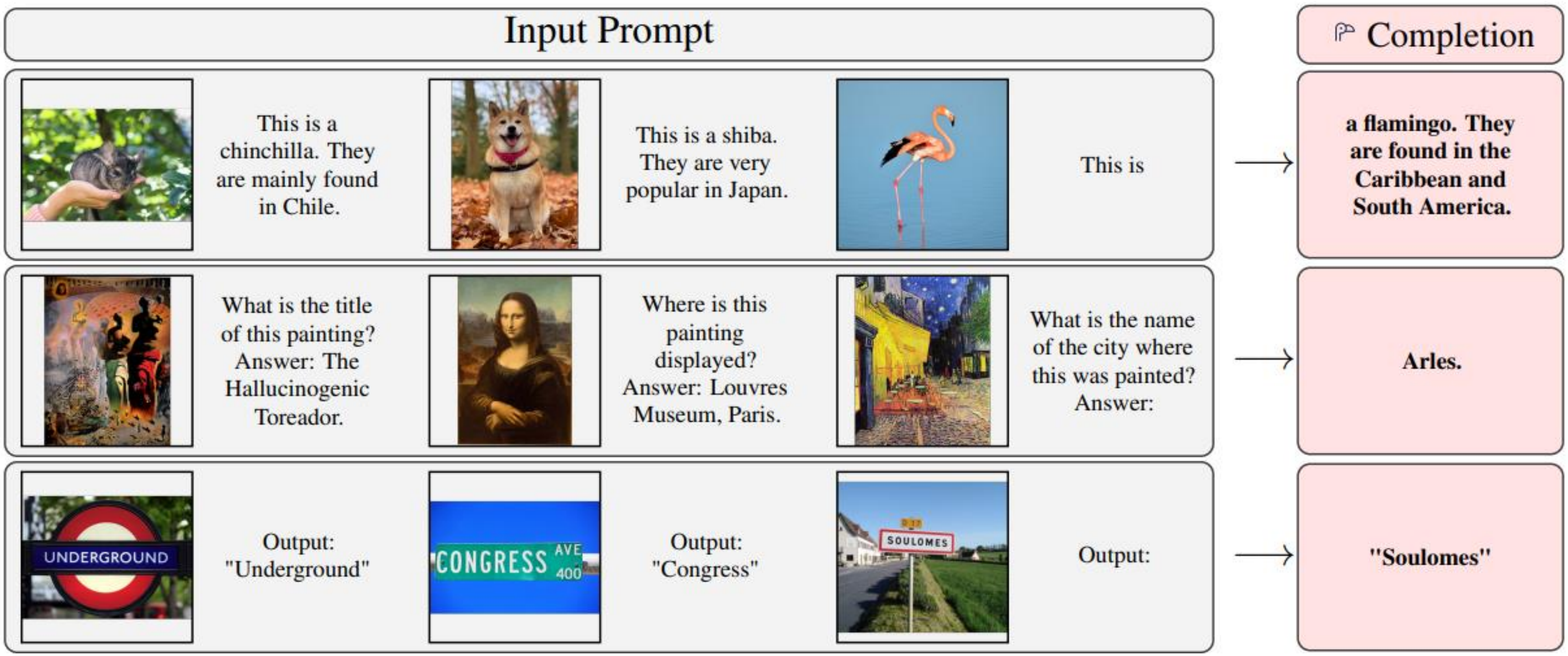
Results



Flamingo: a Visual Language Model for Few-Shot Learning

Alayrac et al., NeurIPS 2022

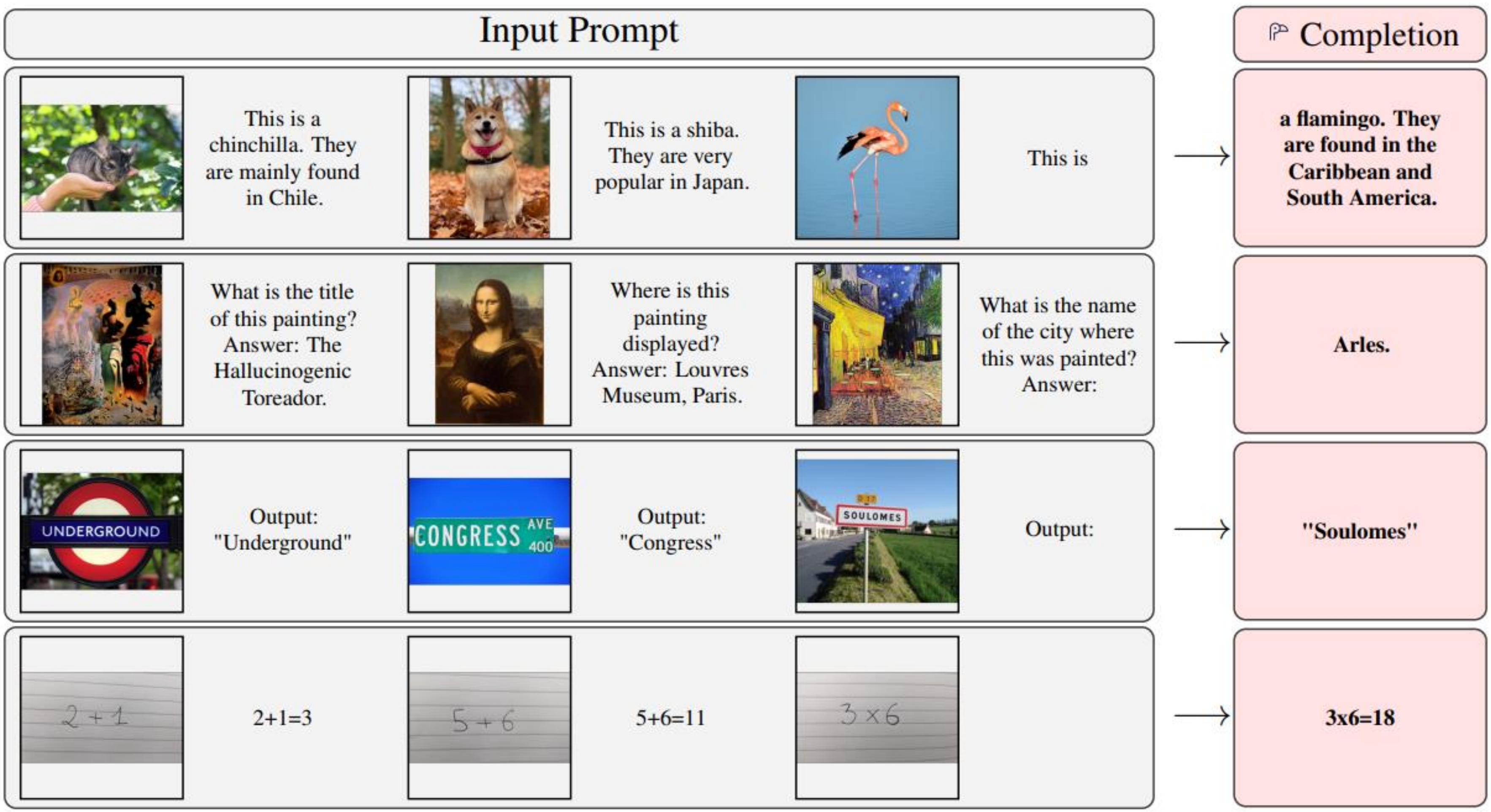
Results



Flamingo: a Visual Language Model for Few-Shot Learning

Alayrac et al., NeurIPS 2022

Results



Flamingo: a Visual Language Model for Few-Shot Learning

Alayrac et al., NeurIPS 2022

Results