Visual recognition: from pixels to machines that see, reason and act

Josef Šivic

CZECH TECHNICAL UNIVERSITY IN PRAGUE

ellis unit | PRAGUE

The research domain: computer vision

… extracting information from images

What human brain sees

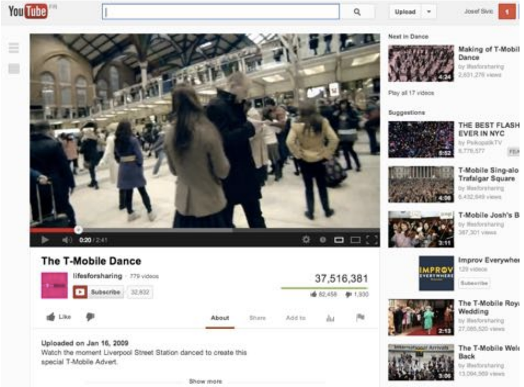
37	43	6	30	36	36	22	48	33	28	26	19	20	14	28	32	27	28	30	38	41	92	26	37	32	28	29	33	162	159	160	159	159	159	149	151	157	61	51	40
65	64	69	48	59	52	59	59	21	47	44	7	55	29	32	67	49	49	45	42	41	108	52	62	60	64	67	81	247	253	254	253	253	253	251	253	253	212	60	87
59	36	73	61	49	64	17	66	58	54	50	44	44	36	27	67	81	4	10	42	42	90	52	79	134	72	88	78	247	251	252	251	251	251	251	251	253	246	96	162
65	70	63	68	56	70	46	52	7	54	43	40	40	31	18	59	59	16	105	36	37	105	58	70	139	75	95	93	244	254	254	252	252	252	253	253	253	253	252	251
66	71	16	14	2	56	47	51	43	49	53	36	47	61	13	16	35	3	6	39	40	106	51	30	53	35	87	87	249	254	243	252	252	252	253	253	253	253	253	255
61	35	64	4	18	55	55	49	19	45	56	52	19	35	20	4	7	62	5	52	39	101	50	38	30	53	78	56	150	252	235	252	243	249	255	255	247	245	252	248
65	67	8	7	1	28	47	57	59	16	57	51	38	29	23	50	16	1	6	32	31	97	62	48	65	65	78	65	174	253	249	228	234	215	80	95	83	82	92	87
9	62	27	19	2	17	59	42	42	2	5	52	45	19	27	8	25	23	31	30	32	89	57	57	53	66	77	83	191	216	221	226	119	73	72	68	64	69	82	32
65	61	30	19	6	35	57	59	57	5	4	34	44	4	36	29	12	16	30	34	35	97	53	59	56	66	71	89	200	217	228	231	85	53	58	56	58	60	13	3
65	66	36	20	19	27	56	6	59	1	13	47	40	2	39	43	27	3	2	19	32	89	83	66	57	62	55	92	133	222	234	255	81	47	51	48	51	22	1	4
39	2	19	26	21	21	59	52	51	8	22	7	43	4	31	45	31	3	26	31	36	82	68	71	52	50	62	69	186	247	240	142	76	39	44	44	46	28	5	7
59	53	35	36	22	16	3	33	61	9	34	47	48	4	13	20	11	13	28	21	27	154	7	25	4	31	16	32	114	253	242	93	76	43	43	45	42	12	14	39
67	54	36	36	13	21	55	55	38	9	24	20	15	7	18	25	35	11	33	46	48	65	46	39	25	24	19	40	164	194	246	94	78	50	49	47	46	12	14	1
66	90	35	43	29	15	48	58	54	6	40	41	45	6	30	54	37	8	13	50	19	44	57	61	52	23	23	44	163	206	239	101	83	46	97	128	122	45	70	41
66	51	41	38	15	42	64	70	62	16	40	59	33	6	11	9	15	15	25	24	31	42	30	39	43	26	30	15	201	205	221	80	74	39	102	120	130	36	47	33
56	62	37	38	9	50	68	64	64	12	43	42	38	3	22	43	28	20	25	19	22	28	27	42	43	11	24	34	63	213	219	55	54	49	46	108	104	78	69	98
81	54	37	32	81	26	53	69	56	16	42	34	45	15	37	48	37	16	36	0	26	26	8	10	45	10	34	27	97	12	78	7	42	57	192	78	105	57	99	52
75	54	37	33	18	34	52	50	56	16	42	28	39	8	31	28	24	9	21	48	49	69	28	40	69	60	1	21	187	111	125	74	103	1	1	27	80	39	33	44
14	9	63	71	75	12	40	43	36	3	4	32	23	34	2	46	32	19	30	0	20	16	24	39	1	40	60	20	48	26	44	32	45	47	50	48	44	37	5	1
29	33	34	53	46	6	27	33	32	28	2	0	41	42	5	20	27	2	22	1	11	3	39	67	64	93	104	93	95	34	8	26	48	45	44	42	41	45	36	29
29	28	28	37	46	3	14	23	30	24	2	3	33	24	5	21	25	4	17	103	112	115	109	119	142	62	48	37	78	45	8	35	37	37	36	35	33	31	35	29
22	21	35	32	49	2	7	29	24	20	2	119	177	37	9	24	50	79	81	88	86	19	81	40	35	62	43	46	63	39	6	33	37	37	36	36	35	32	30	26
22	27	32	35	52	2	14	39	9	12	1	115	113	46	45	38	39	38	36	47	51	72	97	78	60	112	59	5	18	9	38	37	36	36	37	36	35	32	30	26
36	45	25	33	58	4	16	36	14	36	48	125	60	36	34	67	31	38	56	70	22	37	36	68	27	136	96	12	15	19	37	36	36	35	37	36	35	32	31	30
31	40	40	39	62	85	31	70	60	61	75	12	11	60	20	15	6	29	34	34	20	27	63	86	48	168	70	13	3	33	36	39	39	39	39	38	37	36	35	32
42	53	67	73	73	60	81	76	79	68	49	45	41	13	27	21	6	28	26	74	38	34	73	59	65	98	3	1	31	39	39	39	39	41	41	39	38	38	37	33
65	60	63	74	77	76	75	75	68	6	61	37	11	3	21	22	29	60	46	52	57	56	93	48	91	1	19	27	45	47	44	37	39	43	44	44	42	41	39	35
91	82	75	75	85	77	72	76	64	38	42	43	44	50	53	53	57	61	72	70	90	90	93	89	42	51	46	45	40	44	51	45	43	45	45	45	45	43	43	41
89	79	82	85	85	82	78	76	83	39	41	43	41	37	53	53	51	81	83	82	88	85	86	107	48	50	41	42	44	46	46	48	47	46	45	45	45	43	42	39
80	79	82	83	83	85	88	90	34	42	36	39	43	45	44	43	40	64	70	73	74	74	85	38	42	53	34	46	46	49	50	49	48	46	46	46	45	43	41	38
82	85	81	84	81	83	89	92	27	33	36	41	33	34	38	40	35	37	47	46	53	53	76	32	37	45	56	53	53	50	50	51	49	48	46	46	45	43	41	38
78	74	79	68	77	77	76	56	177	134	94	11	13	18	27	22	26	27	23	196	137	129	186	23	39	20	48	50	49	52	49	45	45	45	45	45	45	43	41	38
73	73	65	74	72	65	49	33	80	91	73	18	10	7	65	3	4	3	10	151	80	78	71	29	31	2	37	43	44	45	45	47	47	47	46	46	44	42	41	39
78	81	79	79	69	65	70	1	10	15	16	13	16	16	17	15	19	19	18	23	26	33	34	49	38	50	50	51	51	52	51	48	50	50	48	46	44	42	42	40
117	79	81	79	68	62	10	11	42	52	46	0	108	57	104	68	102	119	4	47	47	40	23	6	32	56	53	55	55	55	54	54	52	50	49	49	46	44	42	41
90	93	92	87	24	0	23	11	1	1	1	2	30	43	58	64	68	47	1	1	1	1	2	8	27	52	56	57	57	57	57	56	54	53	50	50	48	45	44	42
89	81	77	19	20	32	31	27	0	1	1	1	1	2	1	1	1	1	1	1	1	1	1	21	11	61	59	59	59	59	57	56	55	54	52	50	48	45	44	42
97	55	32	48	52	42	35	21	5	1	1	1	1	1	1	1	1	1	1	1	1	1	3	1	10	54	57	56	55	55	54	54	52	49	47	47	47	45	45	44

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]

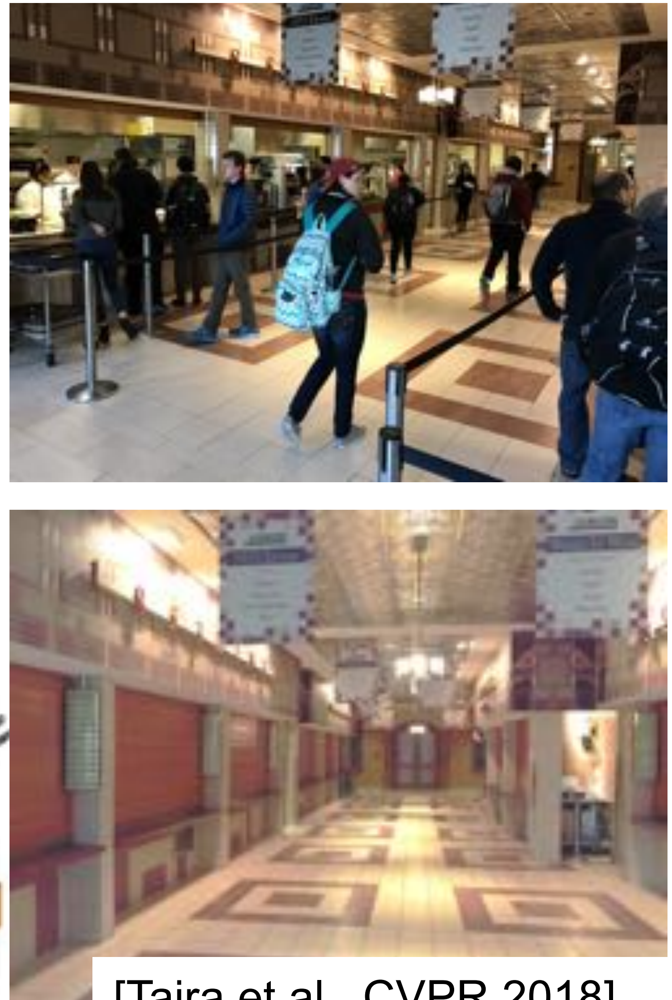
Potential impact: machines that learn from collective visual memory for robotics

Motivation

What if we could automatically learn from this visual data?

Learn to localize and navigate in changing conditions.

[Sattler et al., CVPR 2018]

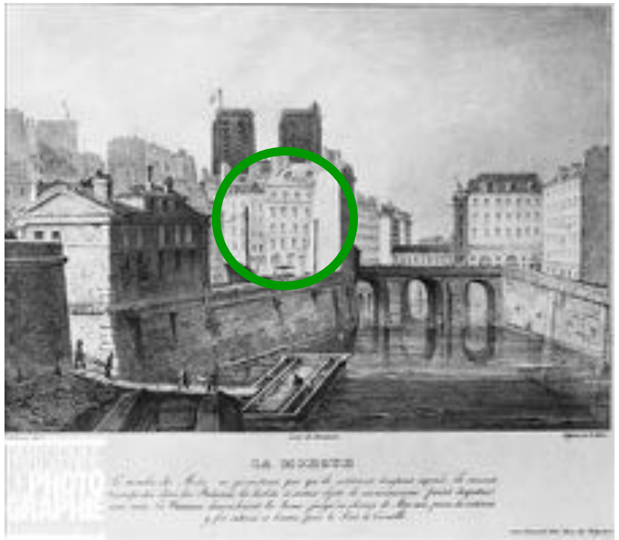


[Taira et al., CVPR 2018]

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

Scientific questions

1. Learning vocabulary of patterns from data

2. Generalization to new conditions and situations

3. Reasoning about visual data

What is the right visual vocabulary?

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)

$f($  $) = +1$

$f($  $) = -1$

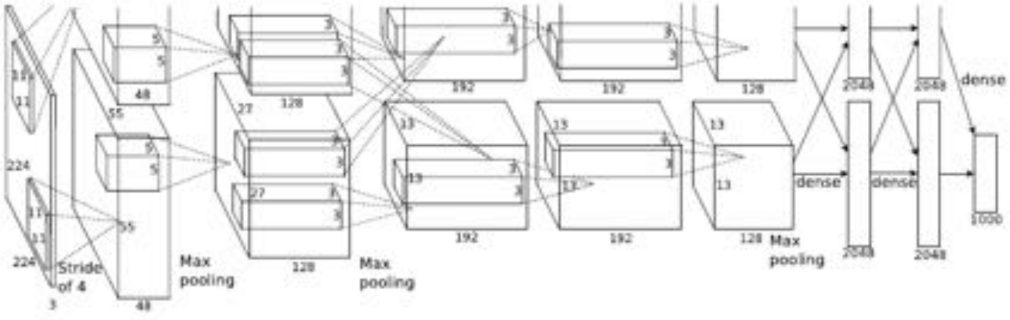
Image classifier

$$\min_{f} \overbrace{\sum_{i=1}^{N} \underbrace{\ell(y_i, f(x_i))}_{\text{Error on one example}}}^{\text{Error (loss) on the all training data}} + \overbrace{\Omega(f)}^{\text{Regularization}}$$

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

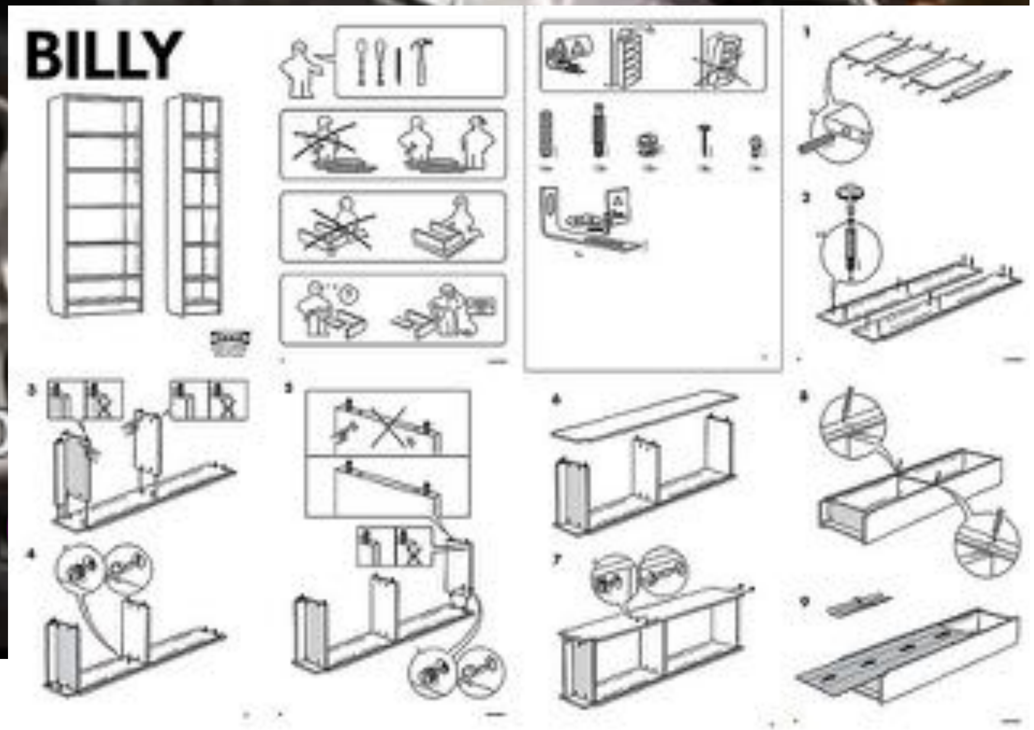
Weakly-supervised learning

Learn from available meta-data :

e.g. video + *text, speech, audio, …*

Learning by interaction with environment
(reinforcement learning)

Examples of meta-data: narrated instructional videos



[Alyarac et al., CVPR 2016]

Learn "vocabulary" of visual patterns from data

Weakly supervised machine learning: [Bach and Harchaoui'08, Xu et al.'04]

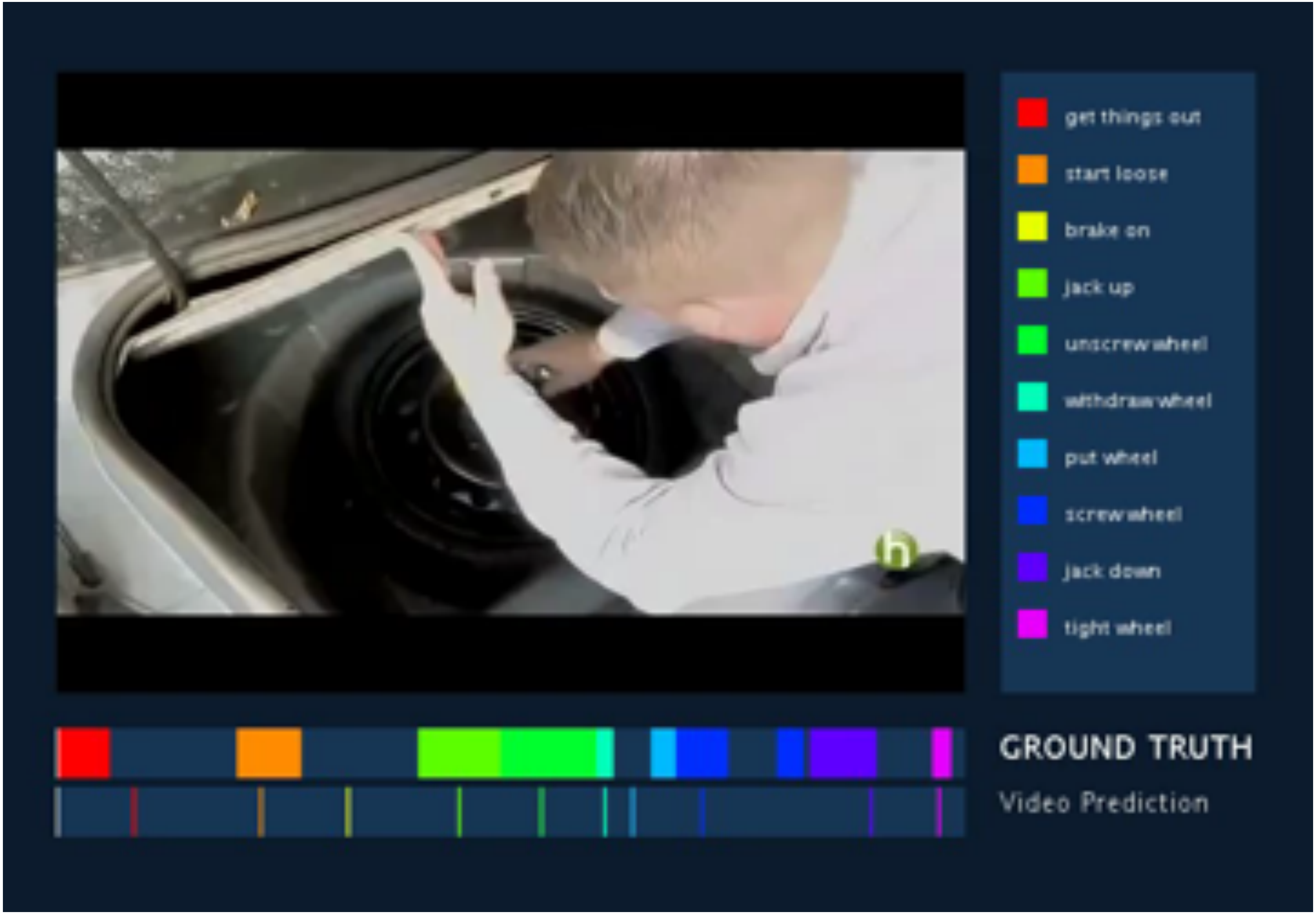
Given a set of inputs x_i and supervisory meta-data y_i, $i = 1,\ldots,N$
learn vocabulary $\hat{z}_i = f(x_i)$ by solving

$$\min_{f,z} \overbrace{\sum_{i=1}^{N} \ell(z_i, f(x_i))}^{\text{Discriminative loss on data}} + \overbrace{\Omega(f)}^{\text{Regularization}}$$

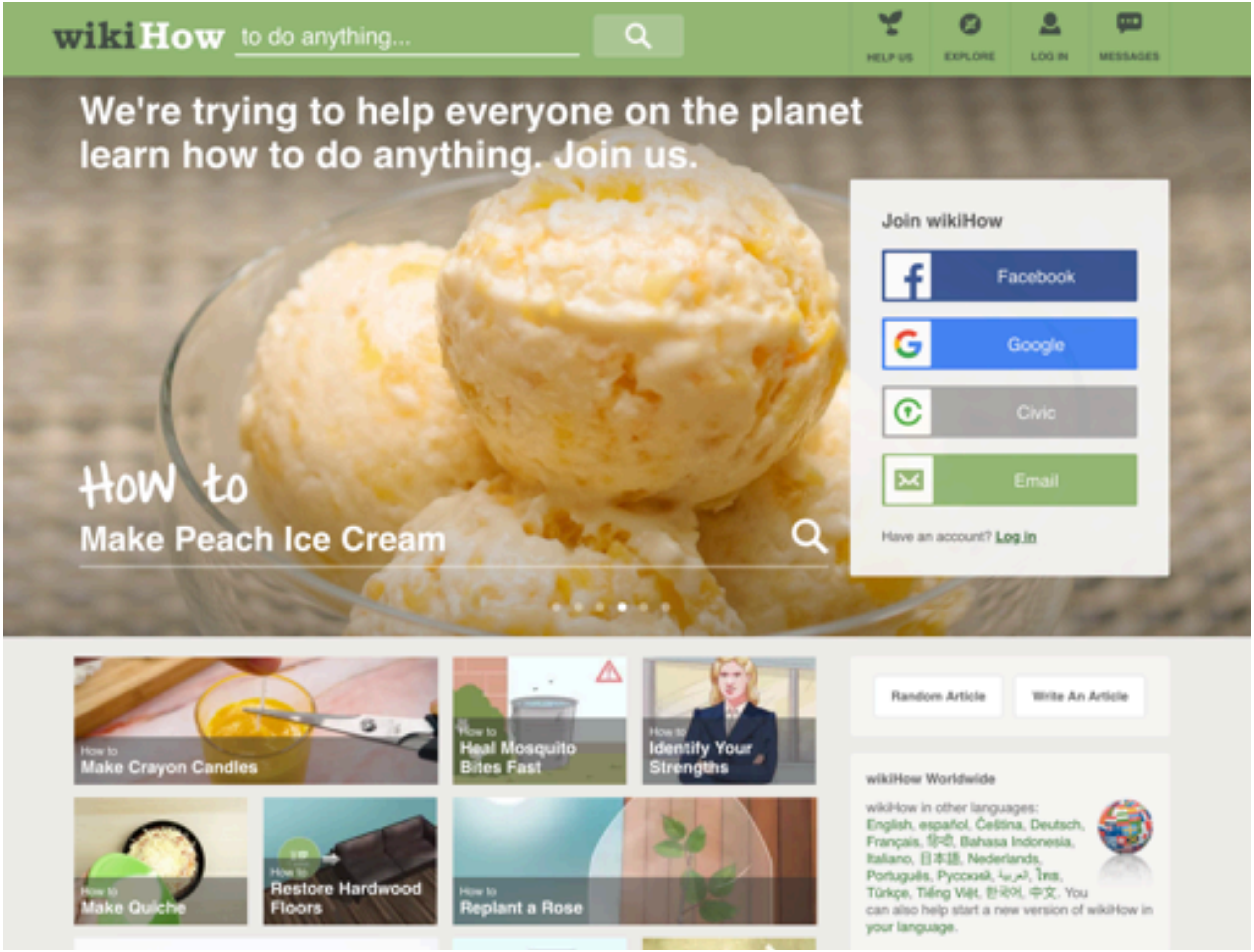
$$\boxed{\underbrace{s.t. \quad Az = y}_{\text{Supervision from available meta-data}}}$$

Scientific challenges:

- What is the appropriate form of constraints to incorporate different types of supervision?
- How to efficiently solve the problem for billions of inputs and 10,000s of patterns?



[Alayrac 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.
Join wikiHow
Facebook
Google
Civic
Email
Have an account? Log in
How to
Make Peach Ice Cream
Random Article
Write An Article
How to
Make Crayon Candles
How to
Heal Mosquito Bites Fast
How to
Identify Your Strengths
wikiHow Worldwide
wikiHow in other languages:
How to
Make Quiche
How to
Restore Hardwood Floors
How to
Replant a Rose
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

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

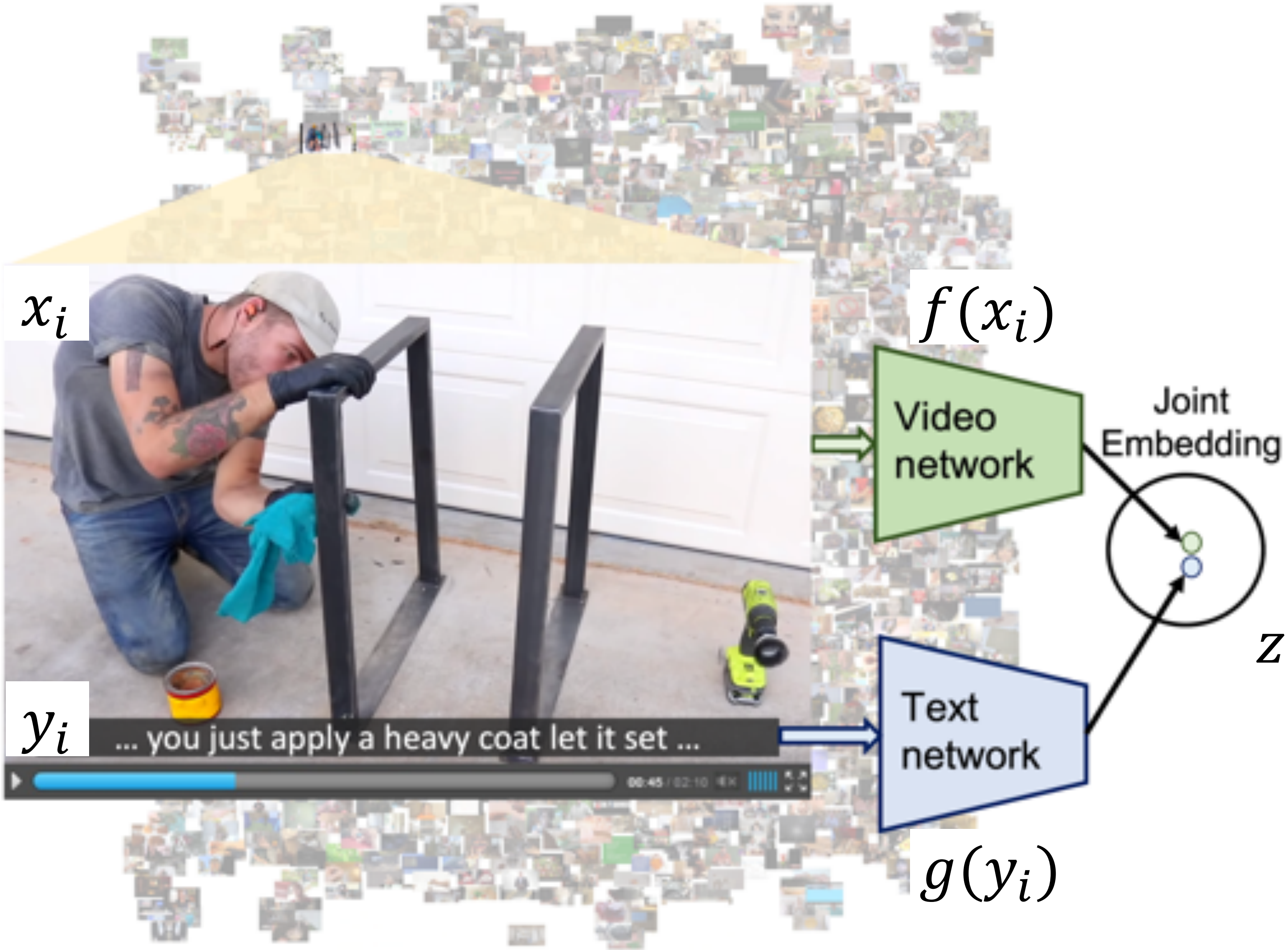
$$\boxed{\text{s.t.} \quad \underbrace{z_i = g(y_i)}_{\text{Supervision from available meta-data}}}$$

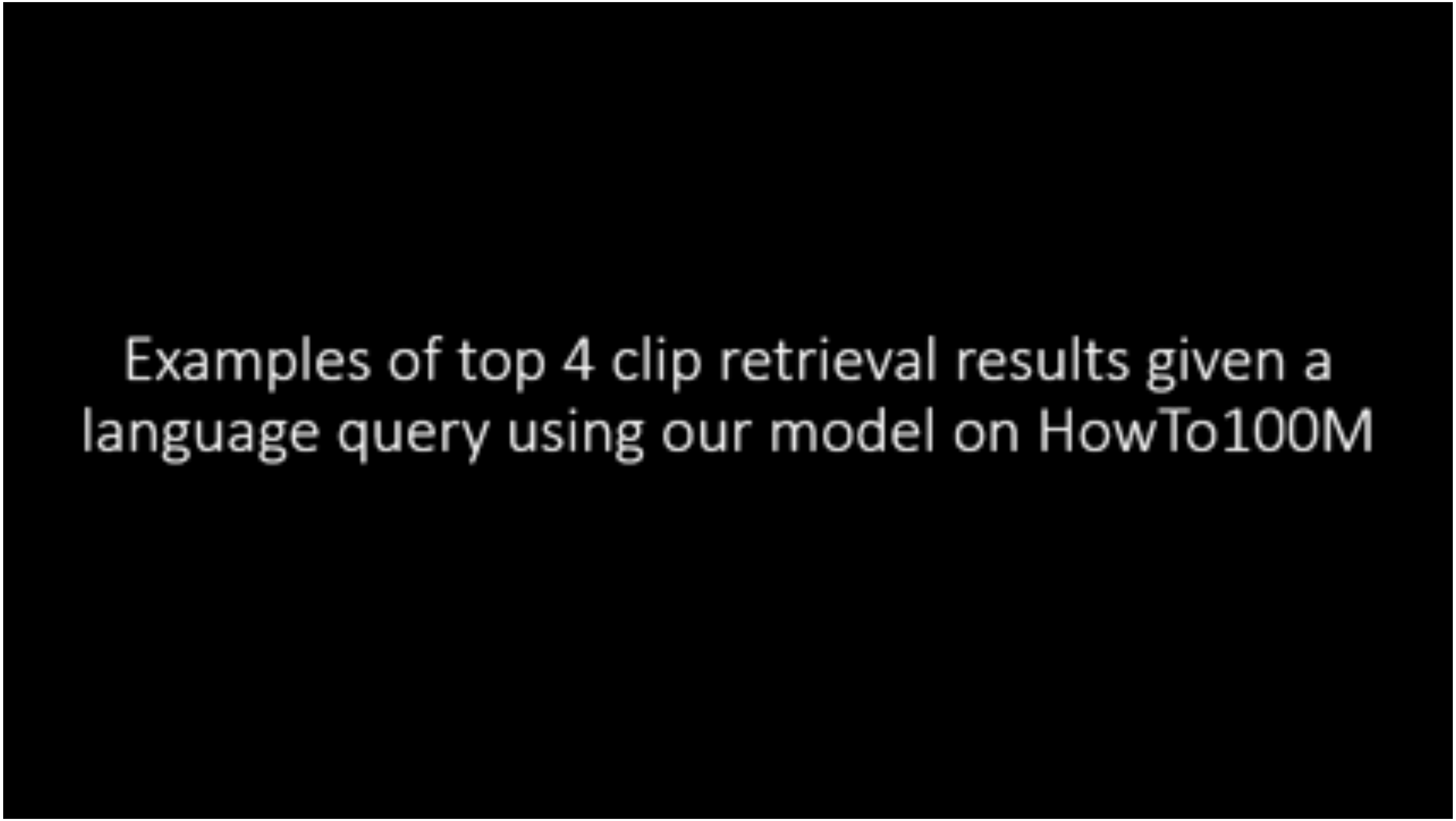
Scientific challenges:

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

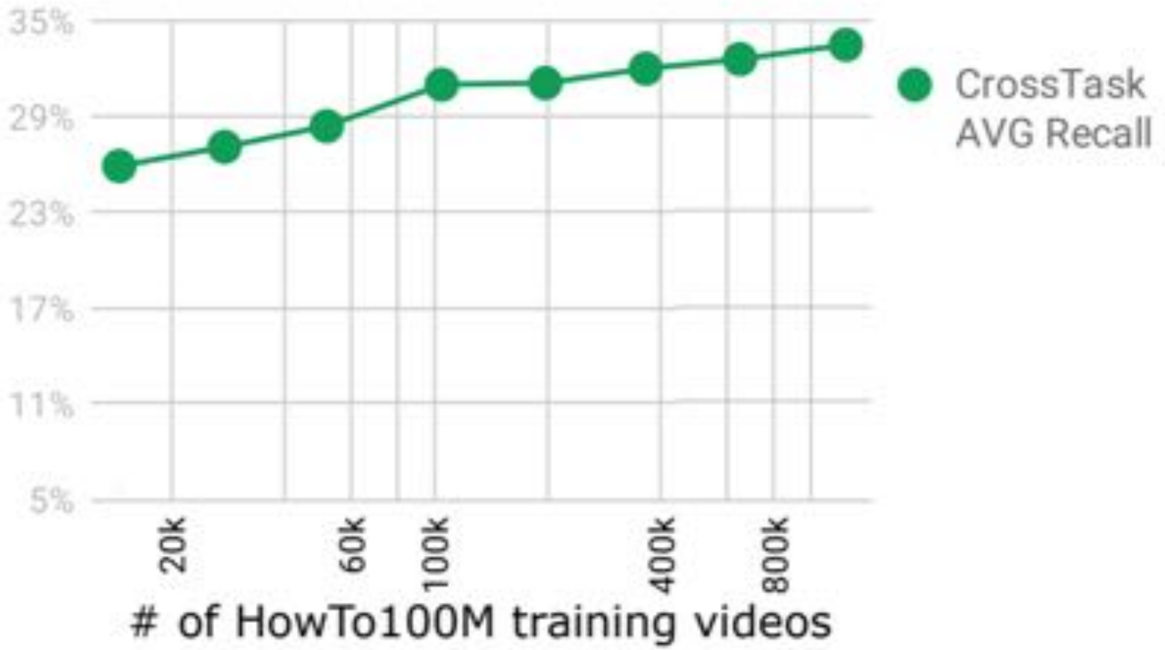
[Gong et al., 2013; Mikolov et al., 2013; Weston et al., 2011; Frome et al., 2013]

Learn joint text-video embeddings from instructional videos



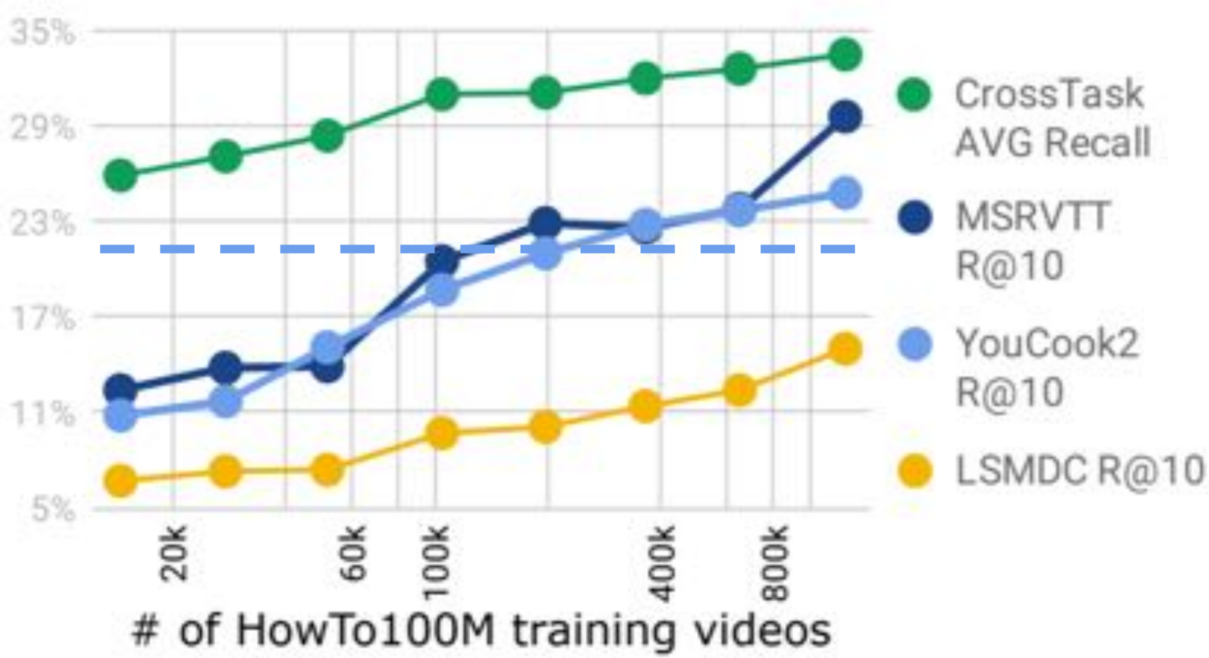

Examples of top 4 clip retrieval results given a language query using our model on HowTo100M

Results: Text-to-video retrieval



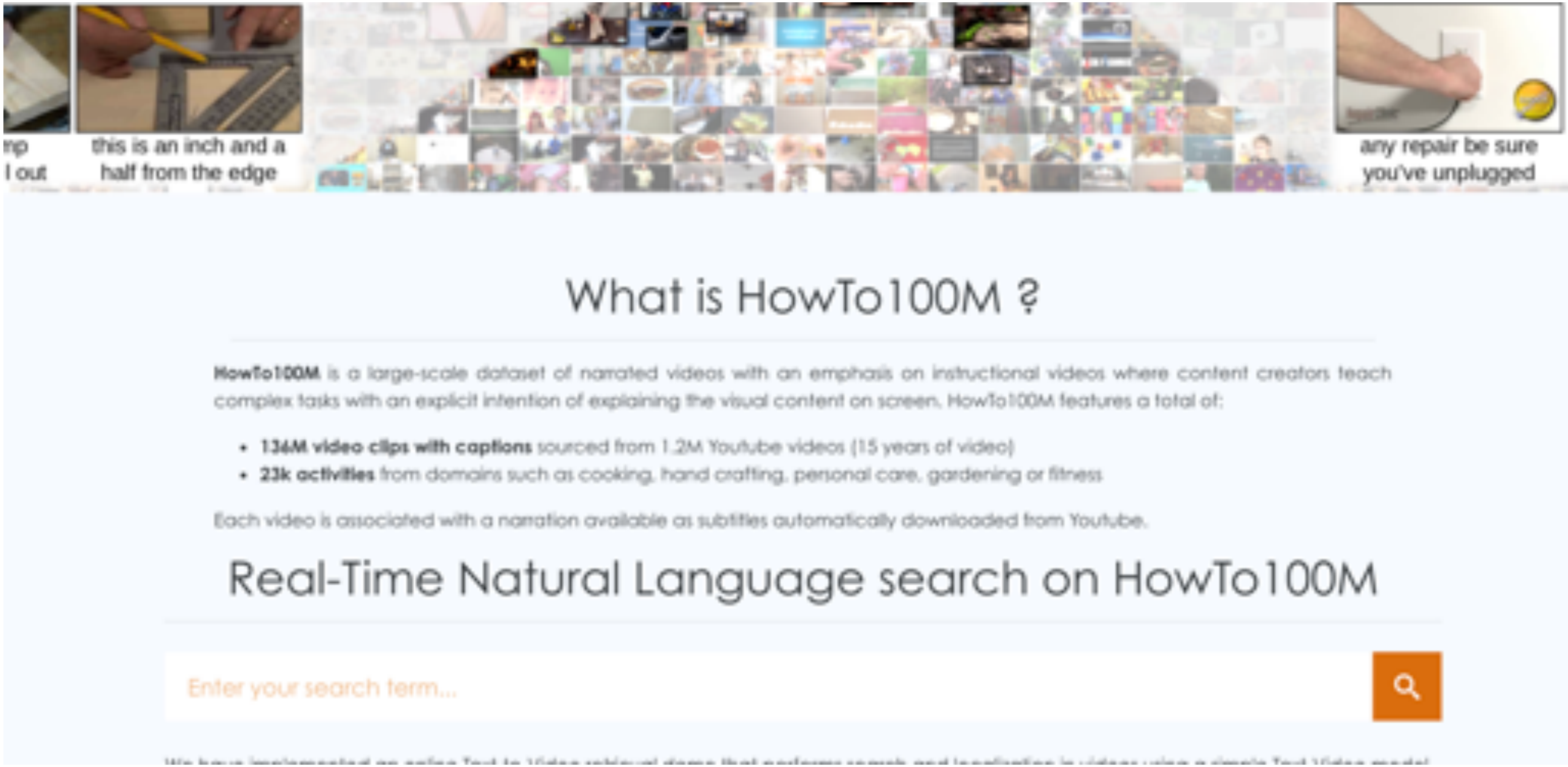
Results: Text-to-video retrieval

Fully supervised with manual annotations



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

Scientific questions

1. Learning vocabulary of patterns from data
2. Generalization to new conditions and situations
3. Reasoning about visual data

How to generalize to new conditions and situations?

Problem: Large image variation due to viewpoint, scale, illumination, occlusion, intra-class variation, …

Different ways to perform the same action

Different viewpoint, occlusion, intra-class variation, …

Scientific challenge: What is the appropriate form of f(x)

$$\min_{f,z} \sum_{i=1}^{N} \ell(z_i, \overbrace{\boxed{f(x_i)}}) \quad + \quad \Omega(f)$$

$$s.t. \quad Az = y$$

- to capture the image variation, and
- can be learnt from few training examples?

Image representations based on convolutional neural networks (CNNs)

Multi-layer nested representation

$$\overbrace{z}^{\text{Output representation}} = \underbrace{f^n}_{\text{“Layer” n}}(\ldots \underbrace{f^2}_{\text{“Layer” 2}}(\underbrace{f^1(\overbrace{x}^{\text{Input image}})}_{\text{“Layer” 1}})\ldots)$$

[Rosenblatt’57], [Hubel&Wiesel’59], [Fukushima’80], [Rumelhart’86], [LeCun et al.’89], [LeCun et al.’98], [Hinton&Salakhutdinov’06], [Krizhevsky’12], …

Image representations based on convolutional neural networks (CNNs)

Multi-layer nested representation

$$\overbrace{z}^{\text{Output representation}} = \underbrace{f^n}_{\text{“Layer” n}}(\ldots \underbrace{f^2}_{\text{“Layer” 2}}(\underbrace{f^1(\overbrace{x}^{\text{Input image}})}_{\text{“Layer” 1}})\ldots)$$

where each layer has a form:

$$f(x) = \sigma(\underbrace{W}_{}x + \underbrace{b}_{})$$

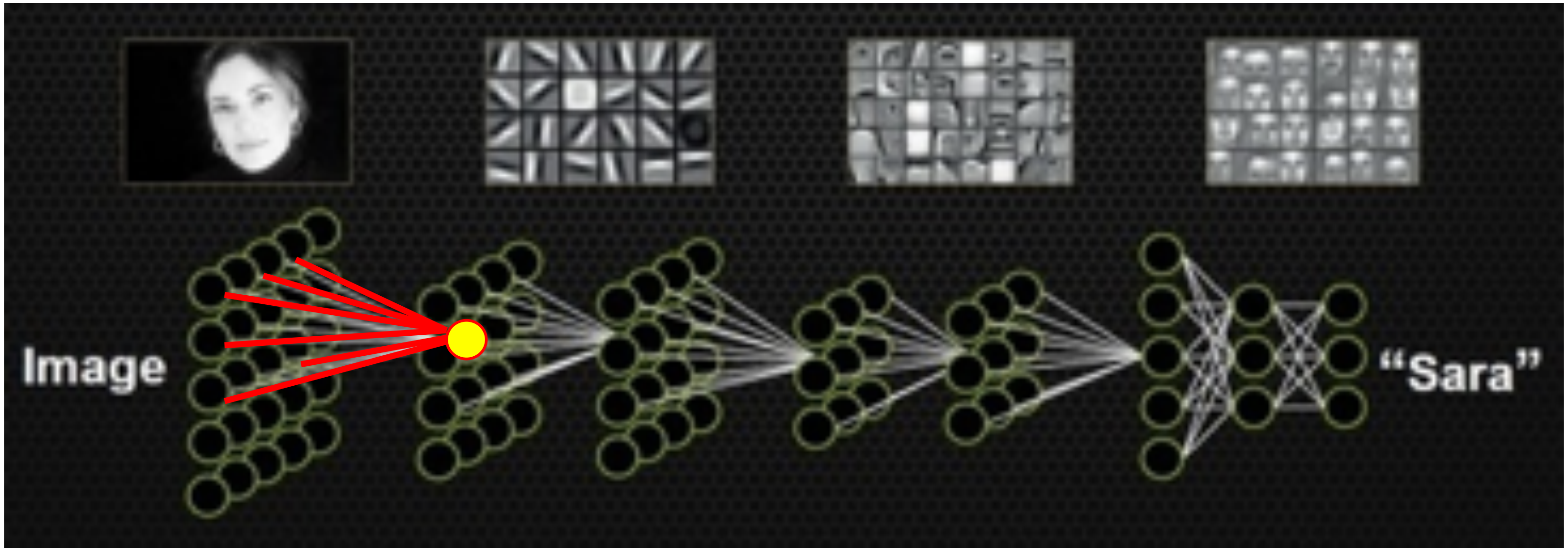
Learnable parameters

[Rosenblatt’57], [Hubel&Wiesel’59], [Fukushima’80], [Rumelhart’86], [LeCun et al.’89], [LeCun et al.’98], [Hinton&Salakhutdinov’06], [Krizhevsky’12], …

Image representations based on convolutional neural networks (CNNs)

Multi-layer nested representation

$$\overbrace{z}^{\text{Output representation}} = \underbrace{f^n}_{\text{"Layer" n}}(\ldots \underbrace{f^2}_{\text{"Layer" 2}}(\underbrace{f^1(\overbrace{x}^{\text{Input image}})}_{\text{"Layer" 1}})\ldots)$$



Source: A. Shivkumar

[Rosenblatt'57], [Hubel&Wiesel'59], [Fukushima'80], [Rumelhart'86], [LeCun et al.'89], [LeCun et al.'98], [Hinton&Salakhutdinov'06], [Krizhevsky'12], …

Image representations based on convolutional neural networks (CNNs)

Multi-layer nested representation

$$\overbrace{z}^{\text{Output representation}} = \underbrace{f^{n}}_{\text{“Layer” n}}(\ldots \underbrace{f^{2}}_{\text{“Layer” 2}}(\underbrace{f^{1}(\overbrace{x}^{\text{Input image}})}_{\text{“Layer” 1}})\ldots)$$

where each layer has a form:

$$f(x) = \sigma(\underbrace{W}_{}x + \underbrace{b}_{})$$

Learnable parameters

The learnt CNN parameters are transferable across tasks.

[Oquab et al. ’13, Oquab et al.‘14],
See also: [Girshick et al.’14, Sermanet et al.’14, Zeiler&Fergus’13, Donahue et al.’13]

Image representations based on convolutional neural networks (CNNs)

Multi-layer nested representation

$$\overbrace{z}^{\text{Output representation}} = \underbrace{f^n}_{\text{"Layer" n}}(\ldots \underbrace{f^2}_{\text{"Layer" 2}}(\underbrace{f^1(\overbrace{x}^{\text{Input image}})}_{\text{"Layer" 1}})\ldots)$$

where each layer has a form:

$$f(x) = \sigma(\underbrace{W}_{}x + \underbrace{b}_{})$$

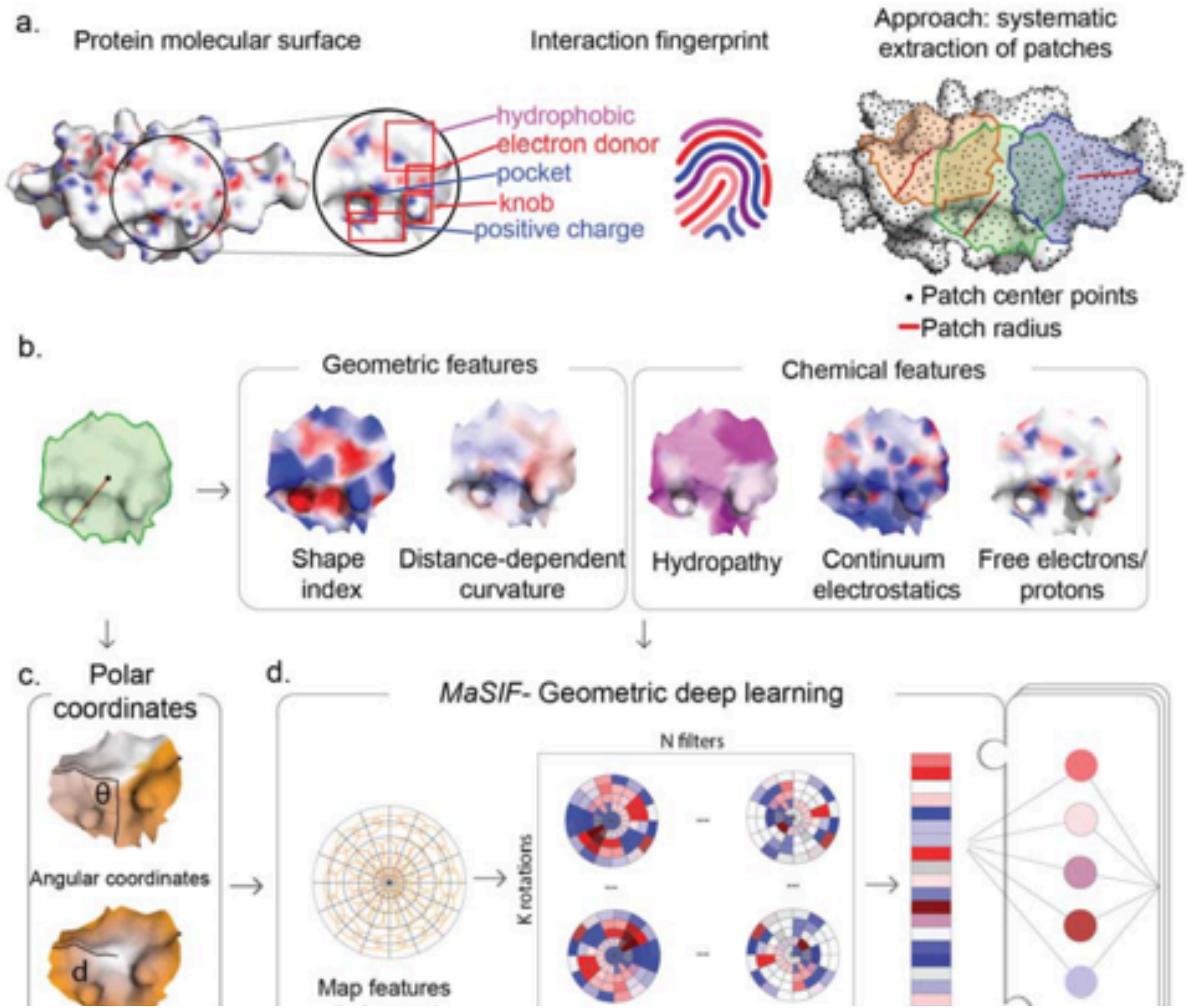
Learnable parameters

But: good for 2D images. Video and 3D objects are still open.

[Oquab et al. ’13, Oquab et al.‘14],
See also: [Girshick et al.’14, Sermanet et al.’14, Zeiler&Fergus’13, Donahue et al.’13]

Input: 3D point cloud
Output: 3D object segmentation

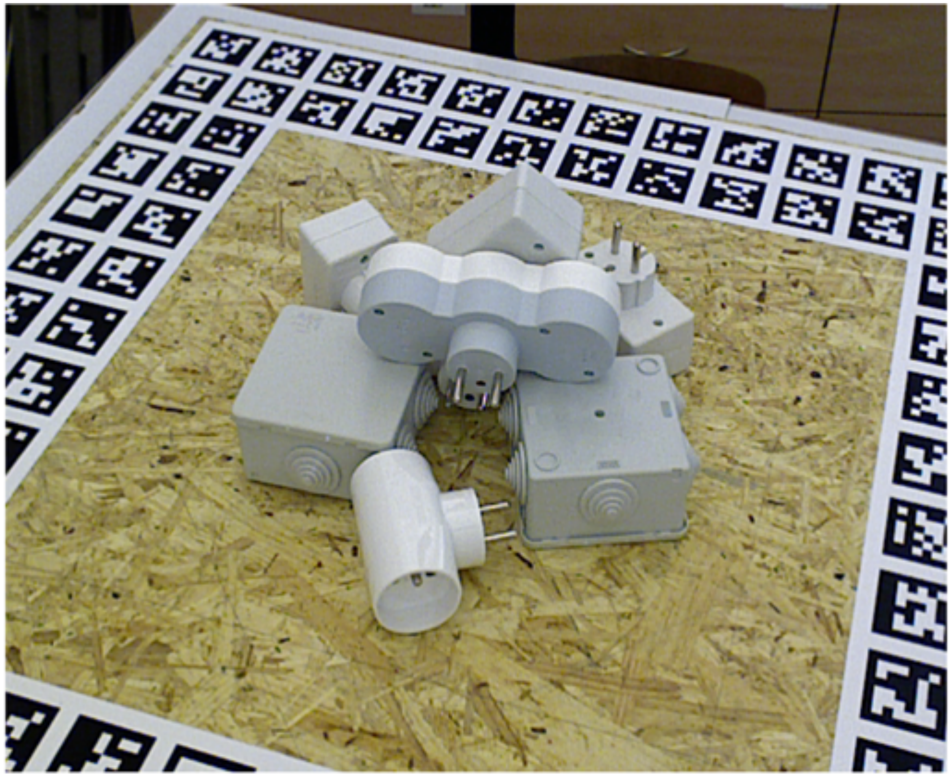
[Engelmann et al., CVPR 2020]



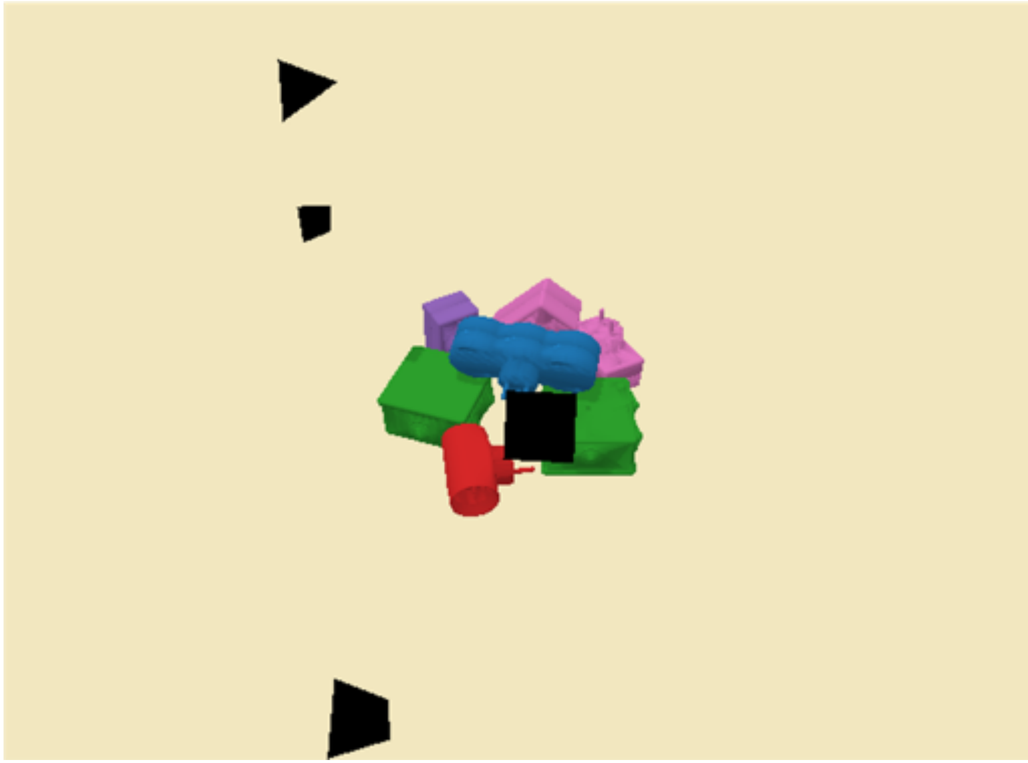
Gainza et al., 2020, Deciphering interaction fingerprints from **protein molecular surfaces using geometric deep learning**, Nature methods.

Object 6D pose estimation

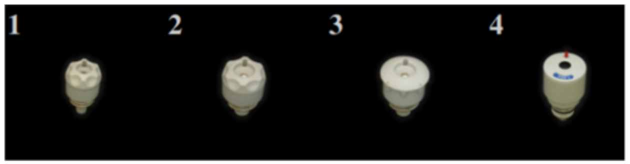
Input image(s)

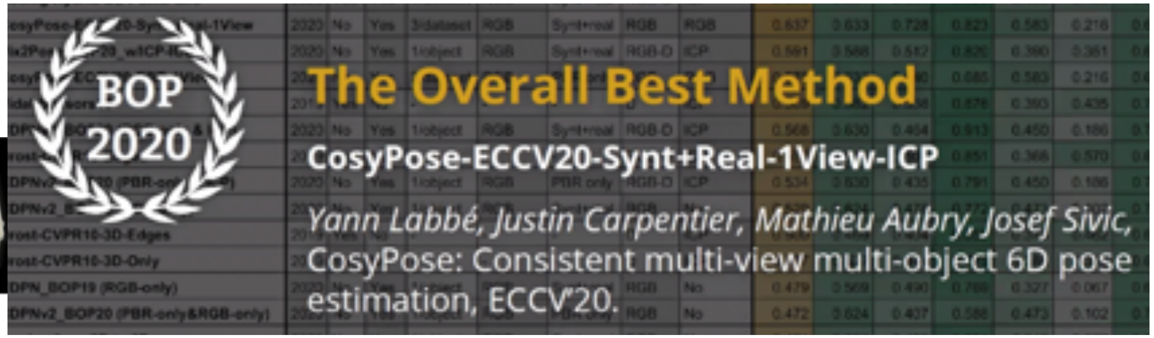


Output 3D scene



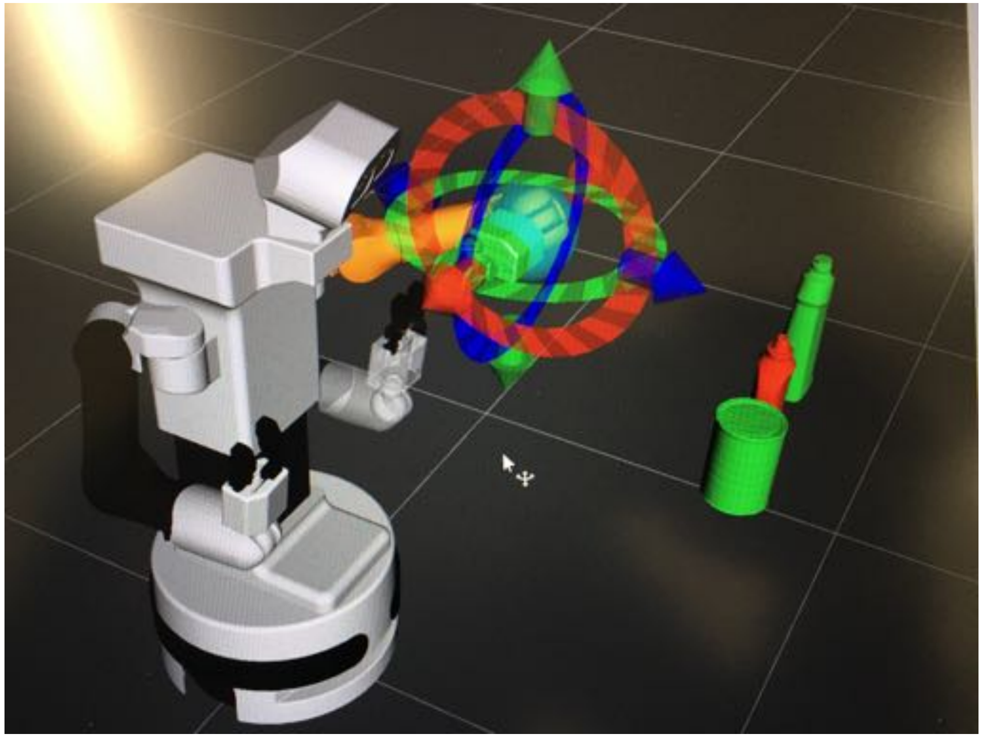
Known 3D models





[Labbe, Carpentier, Aubry, Sivic, ECCV 2020] Code: www.di.ens.fr/willow/research/cosypose/

Towards learnable perception – planning – action

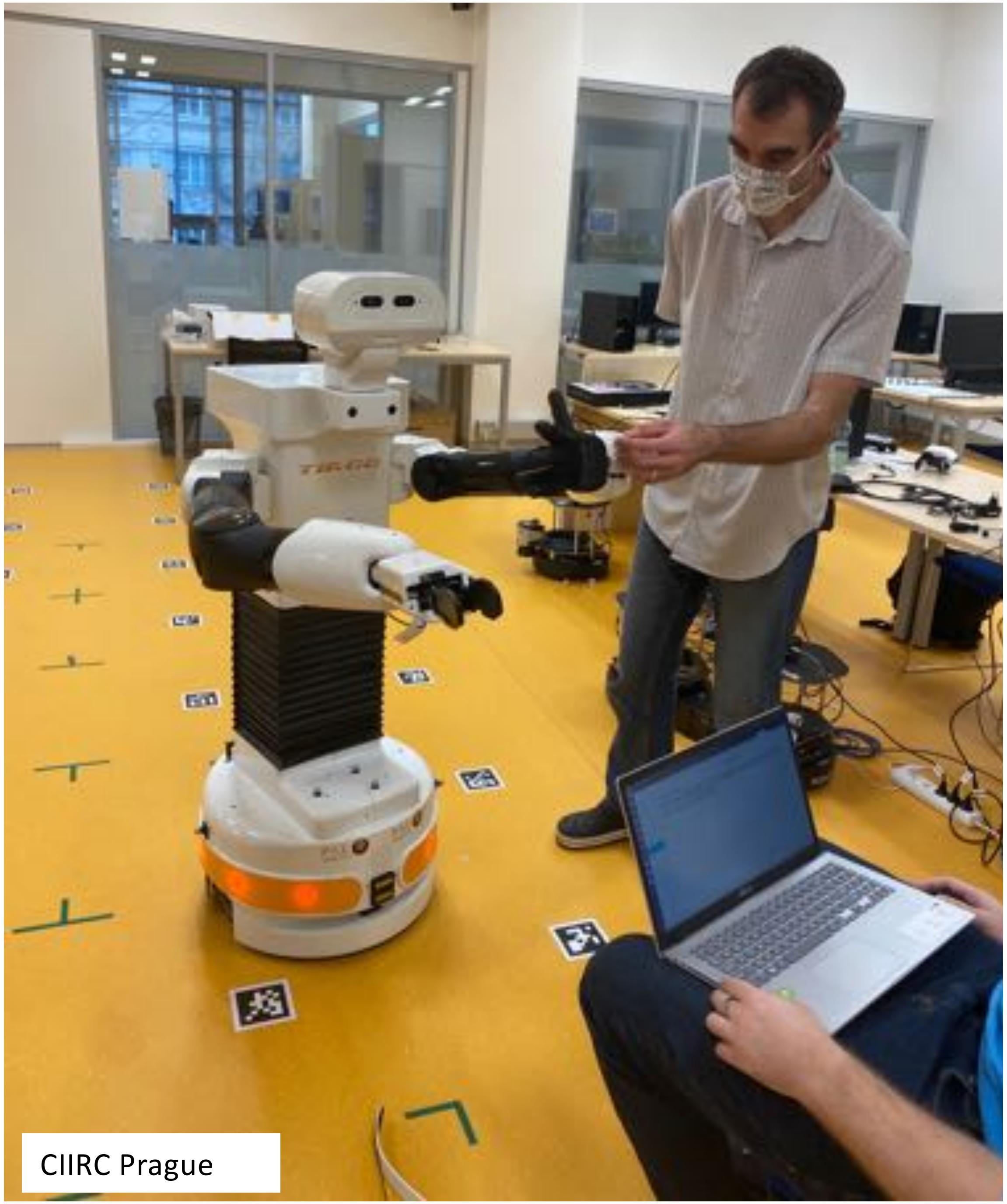


Images by I. Kalevatykh

[Multi-view multi-object 6D pose estimation via robust scene consistency optimization
Y. Labbé, J. Carpentier, M. Aubry, J.Sivic, ECCV 2020]

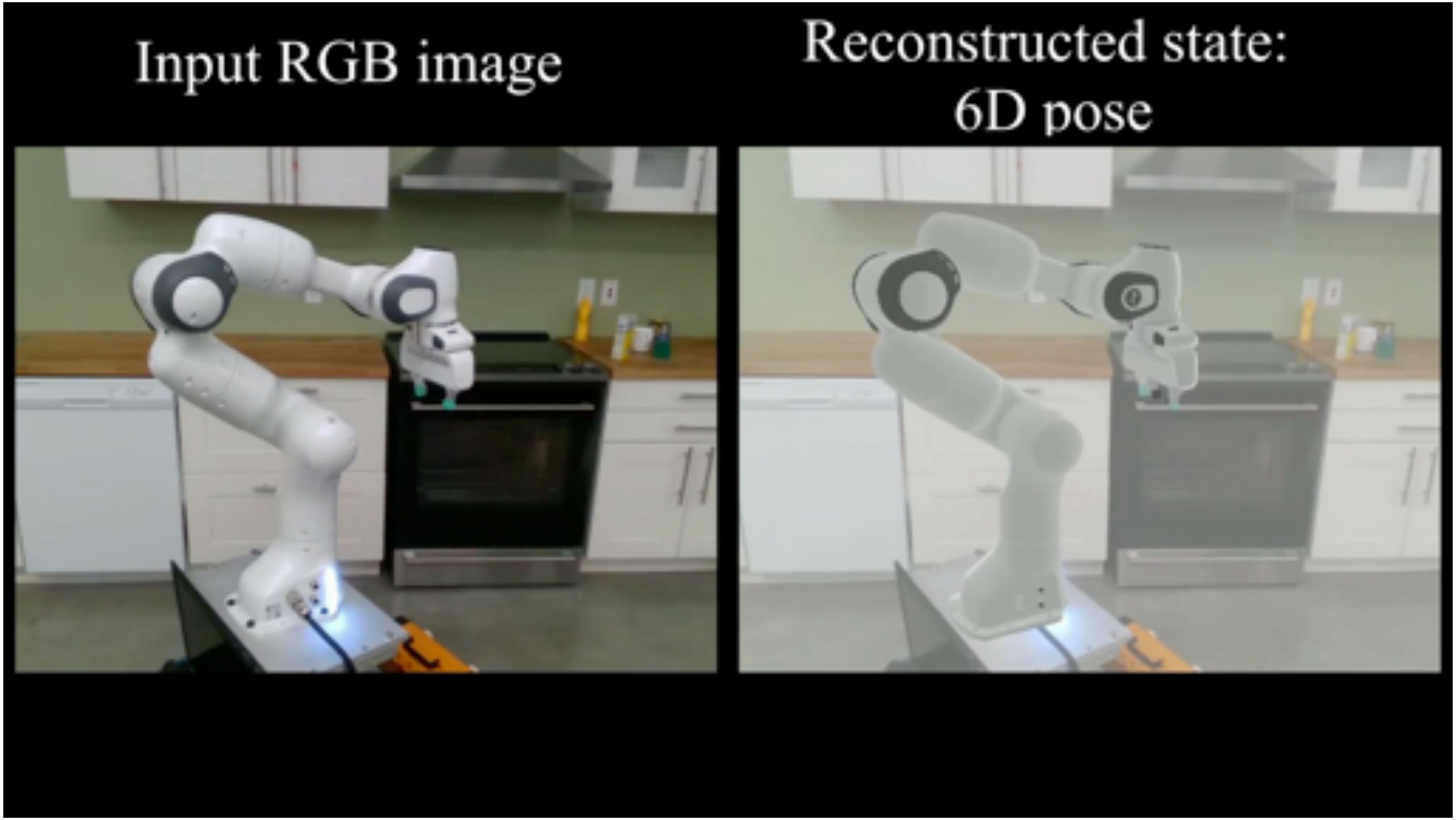
Generalization to different environments

Video by I. Kalevatykh
Inria Paris



CIIRC Prague

6D pose estimation of articulated objects



[Single-view robot pose and joint angle estimation via render&compare
Y. Labbé, J. Carpentier, M. Aubry, J.Sivic, 2020].

Scientific questions

1. Learning vocabulary of patterns from data

2. Generalization to new conditions and situations

3. Reasoning about visual data

What is reasoning about visual data?

Figure 1: Examples from the new GQA dataset for visual reasoning and compositional question answering:
*Is the **bowl** to the right of the **green apple**?*
*What type of **fruit** in the image is **round**?*
*What color is the **fruit** on the right side, red or **green**?*
*Is there any **milk** in the **bowl** to the left of the **apple**?*

[Hudson and Manning, CVPR 2019]
Visualreasoning.net

Recognizing relations between entities is hard

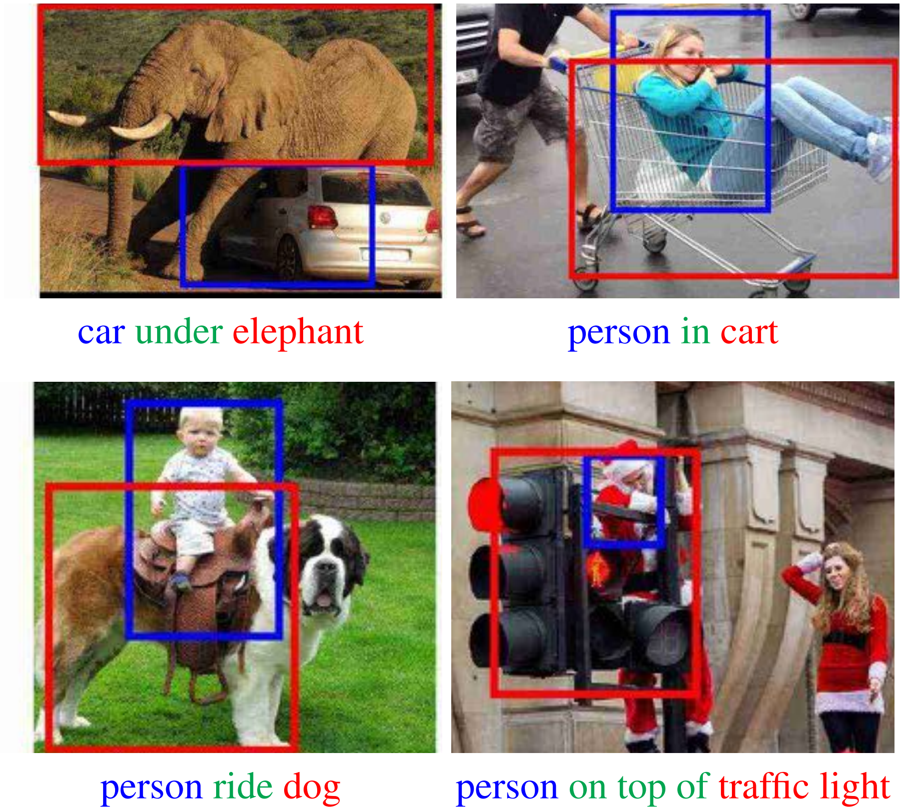
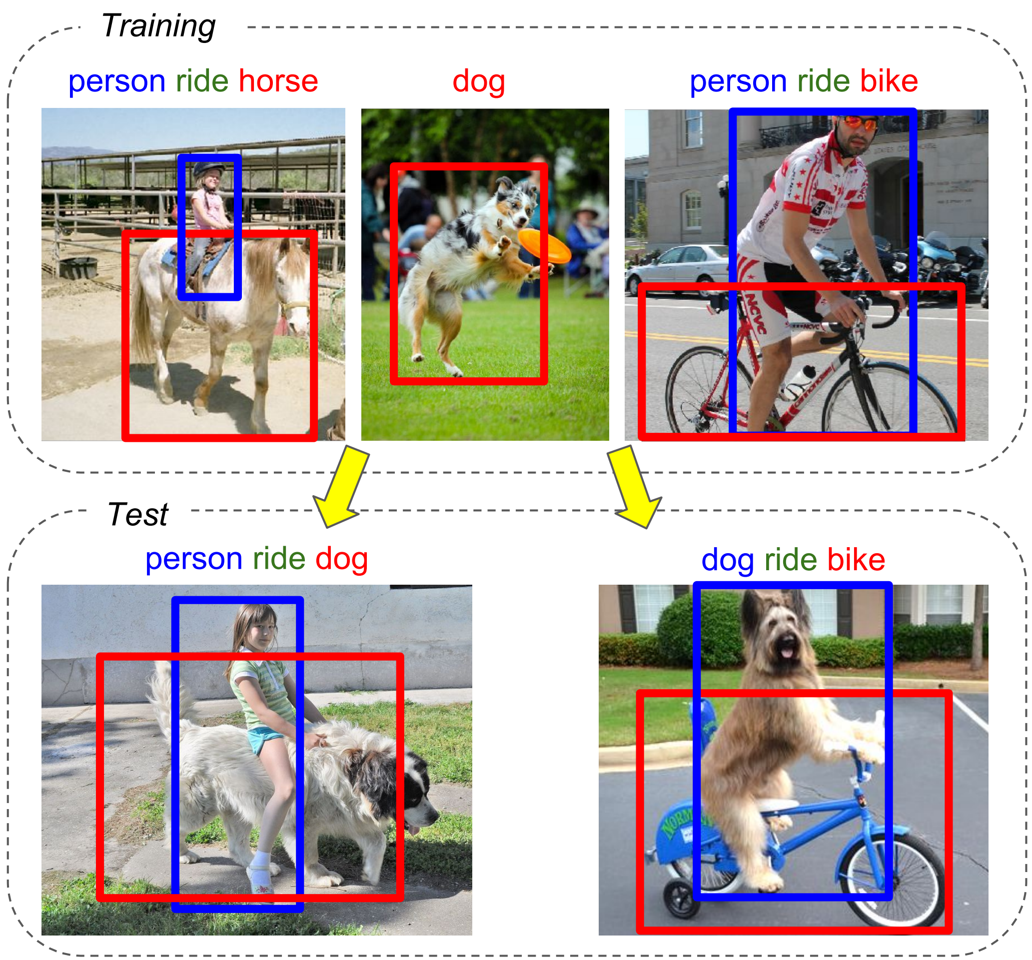


Figure 1: Examples of top retrieved pairs of boxes in UnRel dataset for unusual queries (indicated below each image) with our weakly-supervised model described in 3.2.

[Peyre, Laptev, Schmid, Sivic, ICCV 2017]



[Peyre, Laptev, Schmid, Sivic, ICCV 2019]

Neuro-symboling reasoning?

Differentiable first order logic

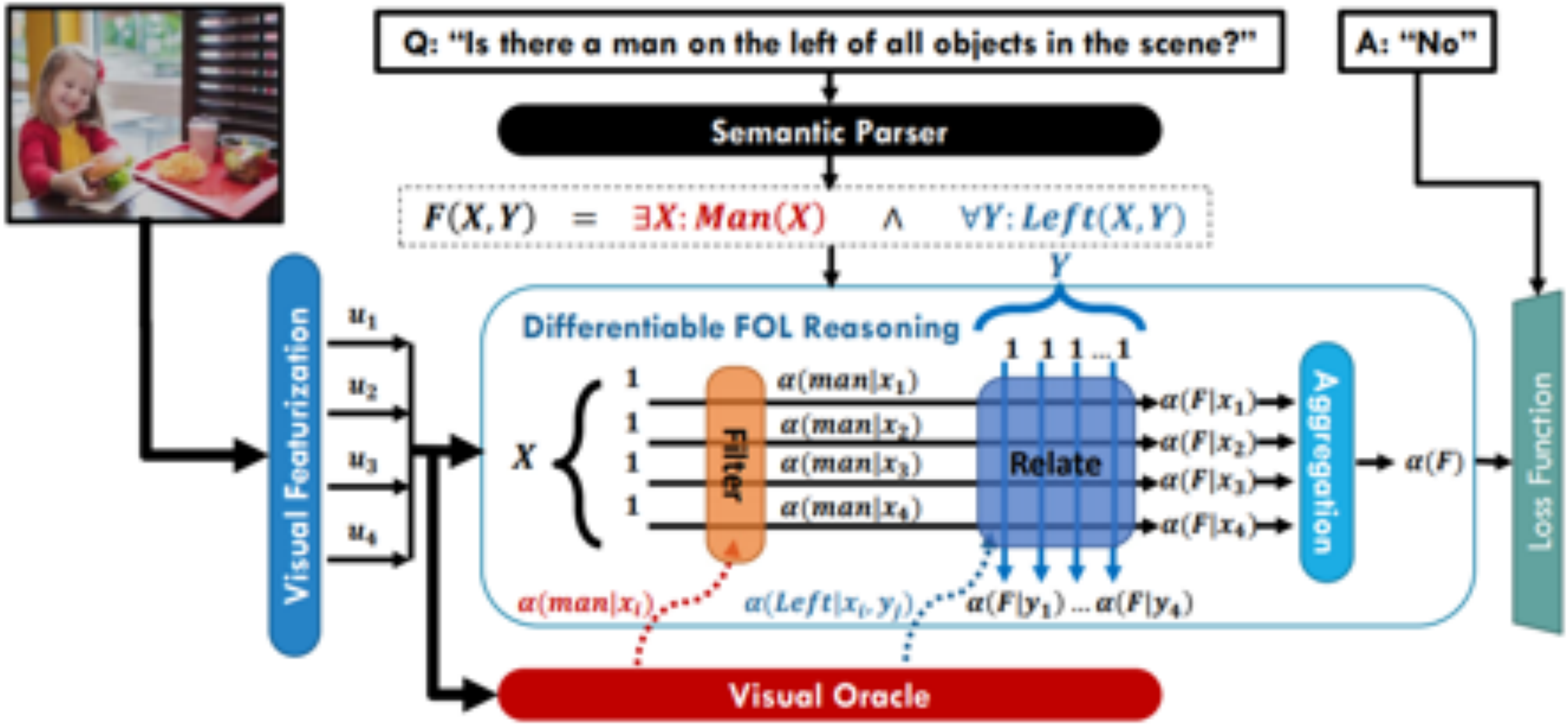


Figure 1. The multi-step question answering process in the ∇-FOL framework, based on differentiable first-order logic.

[Neuro-Symbolic Visual Reasoning: Disentangling "Visual" from "Reasoning"
Saeed Amizadeh, Hamid Palangi, Oleksandr Polozov, Yichen Huang, Kazuhito Koishida, ICML 2020.]

Learn to “reason implicitly” (from lots of data)

Just Ask: Learning to Answer Questions from Millions of Narrated Videos

Antoine Yang[1,2], Antoine Miech[1,2,+], Josef Sivic[3], Ivan Laptev[1,2], Cordelia Schmid[1,2]

[1]ENS [2]Inria Paris [3]CIIRC CTU

https://www.di.ens.fr/willow/research/just-ask/

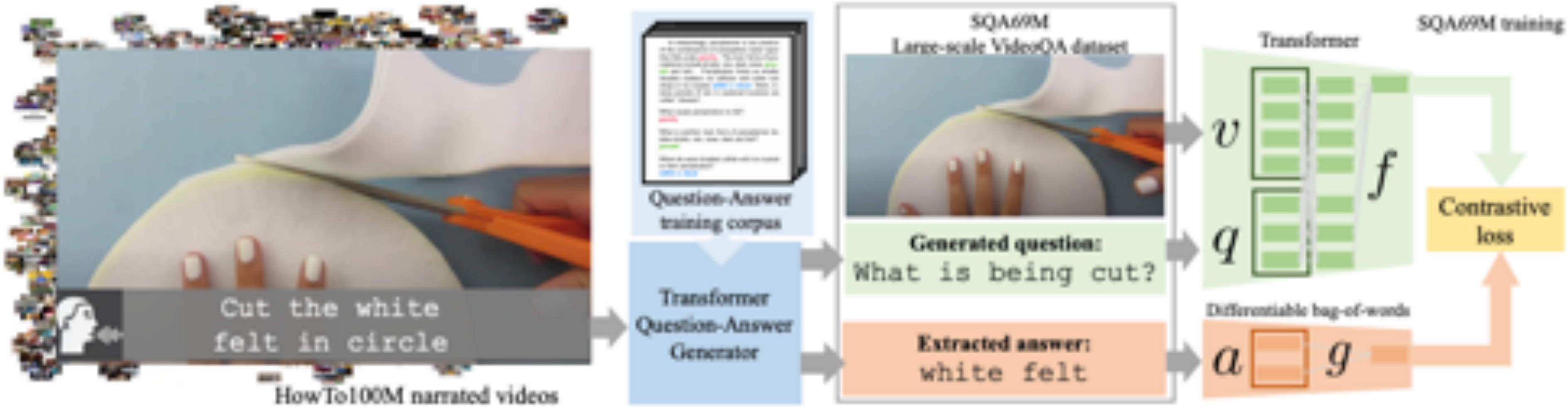


Figure 1: We leverage millions of narrated videos and improve VideoQA with automatic pretraining. We generate question and answer pairs from speech transcripts with a state-of-the-art text-to-text transformer pipeline. Then we use the generated dataset to train a VideoQA model with a contrastive loss *without additional visual annotation*. The pretrained model can then be used for zero-shot or finetuning.

[https://arxiv.org/pdf/2012.00451.pdf]

Learn to “reason implicitly” (from lots of data)

Question: What type of material is the man touching?
GT Answer: wood (5)
VQA-MMT+PT-QA: leather
VQA-MMT+PT-VA: clamps
Ours: wood

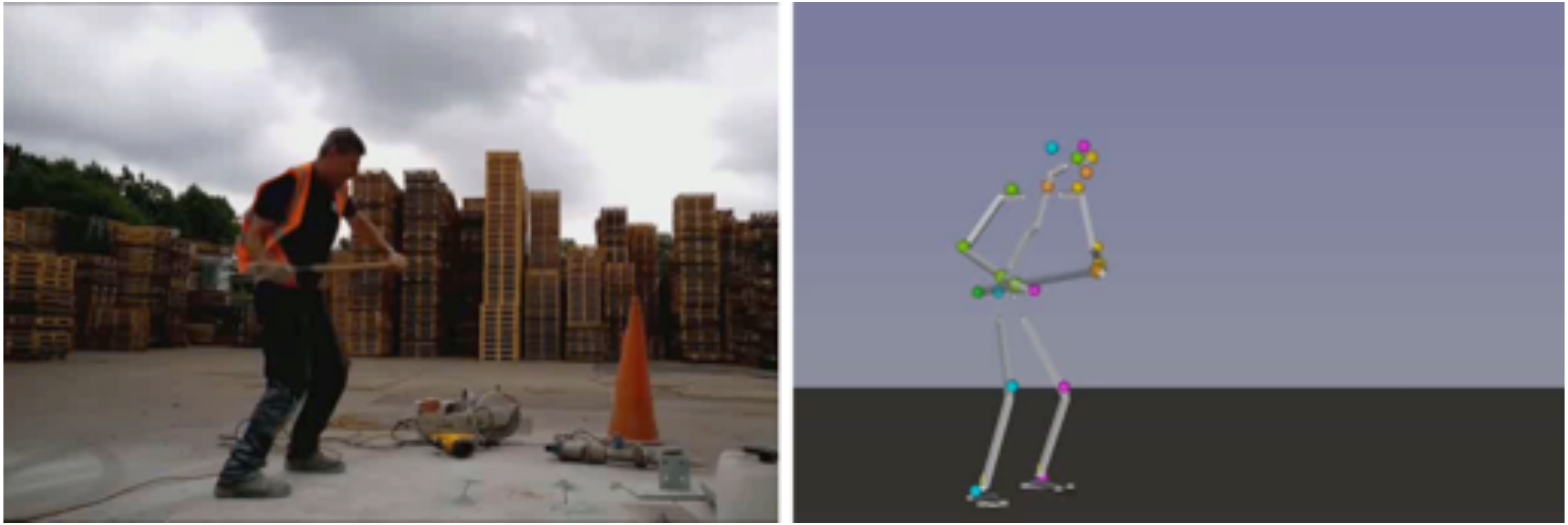
Question: What animal is shown as a cutout?
GT Answer: deer (3), reindeer (2)
VQA-MMT+PT-QA: wolf
VQA-MMT+PT-VA: paintbrush
Ours: reindeer

[https://arxiv.org/pdf/2012.00451.pdf]

Relations are dynamic and in 3D

Input:
- a monocular RGB video

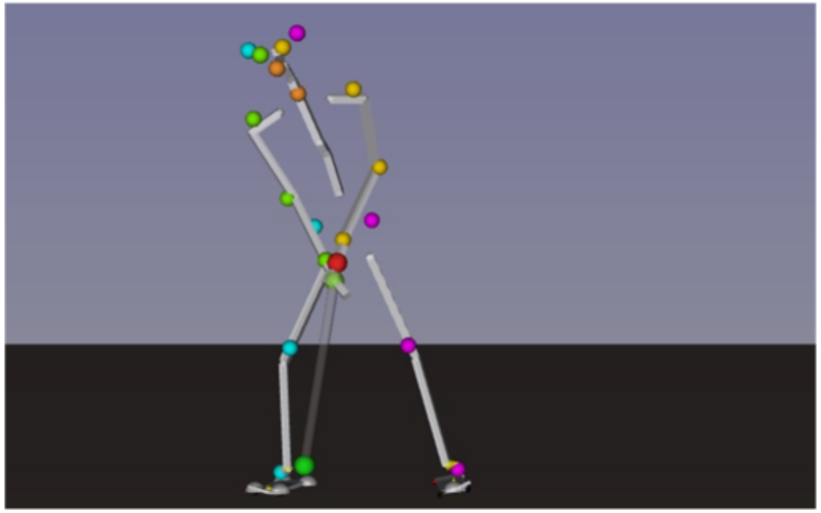
Output:
- Person & object 3D motion trajectories
- Contact positions and contact forces



[Li, Sedlar, Carpentier, Mansard, Laptev, Sivic, CVPR 2019, **best paper finalist**]

Estimation Stage

Problem formulation



$$\underset{\underline{x},\underline{u},\underline{c}}{\text{minimize}} \quad \sum_{e\in\{\mathrm{h},\mathrm{o}\}} \int_0^T l^e(x,u,c)\,\mathrm{d}t, \quad \text{(Objective function)}$$

$$\begin{aligned}\text{subject to} \quad & \kappa(x,c) = 0 \quad \text{(contact motion model)},\\ & \dot{x} = f(x,c,u) \quad \text{(full-body dynamics)},\\ & u \in \mathcal{U} \quad \text{(force model)},\end{aligned}$$

Relations can change objects

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]

Can we learn to reason and plan from data?

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

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

$$\boxed{\underbrace{\text{s.t.} \quad z_i = g(y_i)}_{\text{Supervision from available meta-data}}}$$

Scientific challenges:

- How to incorporate the geometric and physical constraints on the latent space z?
- How to learn such constraints from data?

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

Scientific questions

1. Learning vocabulary of patterns from data
2. Generalization to new conditions and situations
3. Reasoning about visual data
4. Plan and Act on the world. Learn from the interactions

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]

Towards intelligent perception for the real world

Soon: We will see more applications in specific constrained set-ups.

[Microsoft HoloLens]

[Darpa robot challenge]

Long-term: autonomous learning, reasoning and interaction.

Collaboration with other research domains: machine learning, robotics, natural language processing, speech understanding, control, ...

Thank you