Structured Deep Learning for Video Analysis

Fabien Baradel
PhD Candidate

Advisors: Christian Wolf & Julien Mille

June, 29th, 2020

fabienbaradel.github.io

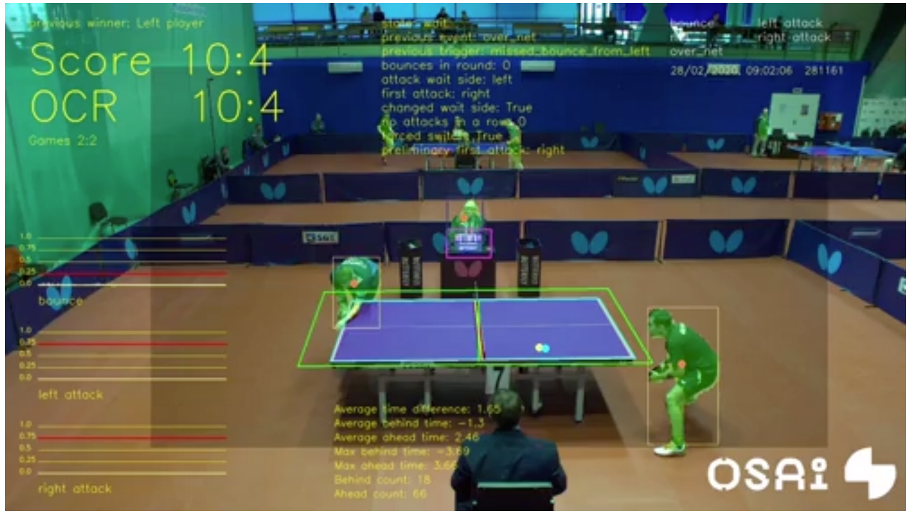
What is video understanding?

Human actions → Sitting on the floor

Entity-level interactions → Grabing a silencer

Temporal reasoning / Causality → The baby starts crying because of the silencer

Why video understanding?



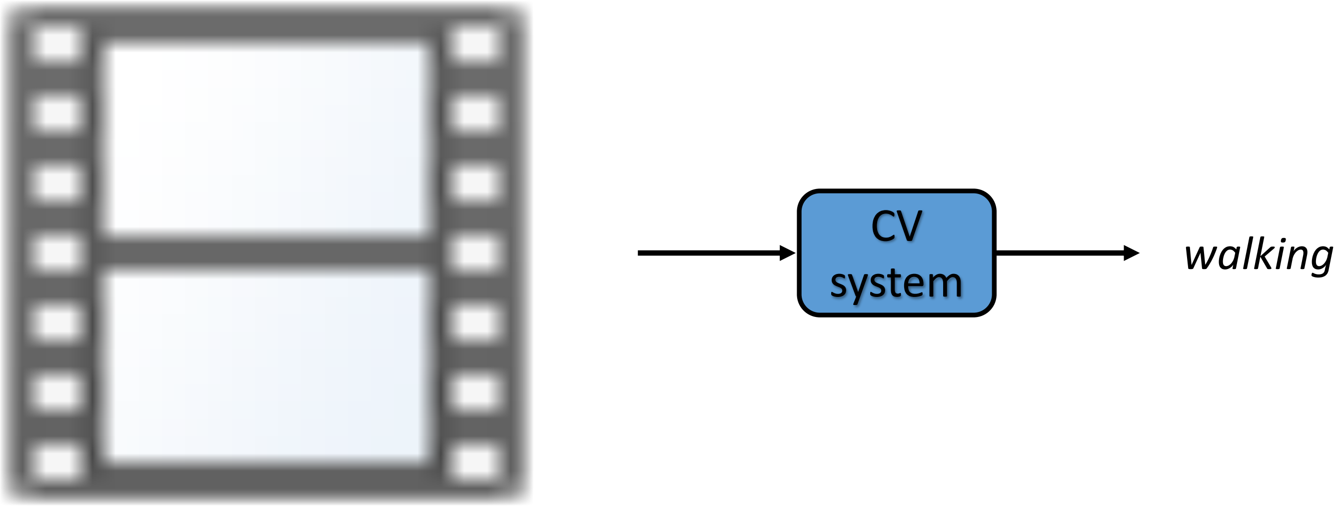
Indexing
Retrieval

Recommendation
Analysis

Human-robot interactions

[TTNET, CVPRW'20]

A video understanding task
Action Recognition

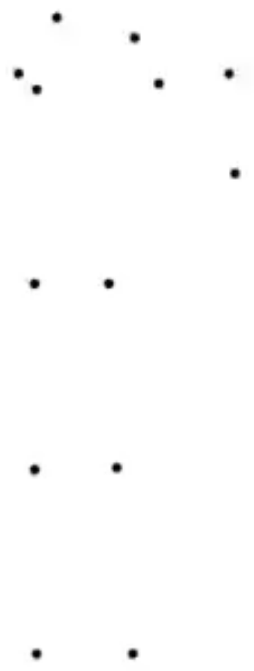


Classification task
Pre-defined labels
Similar to object recognition

[HMDB, Kuehne, ICCV'11]

Human Pose

Walking
Pose is enough



Chopping
What is being chopped?

[Johansson, 1973]

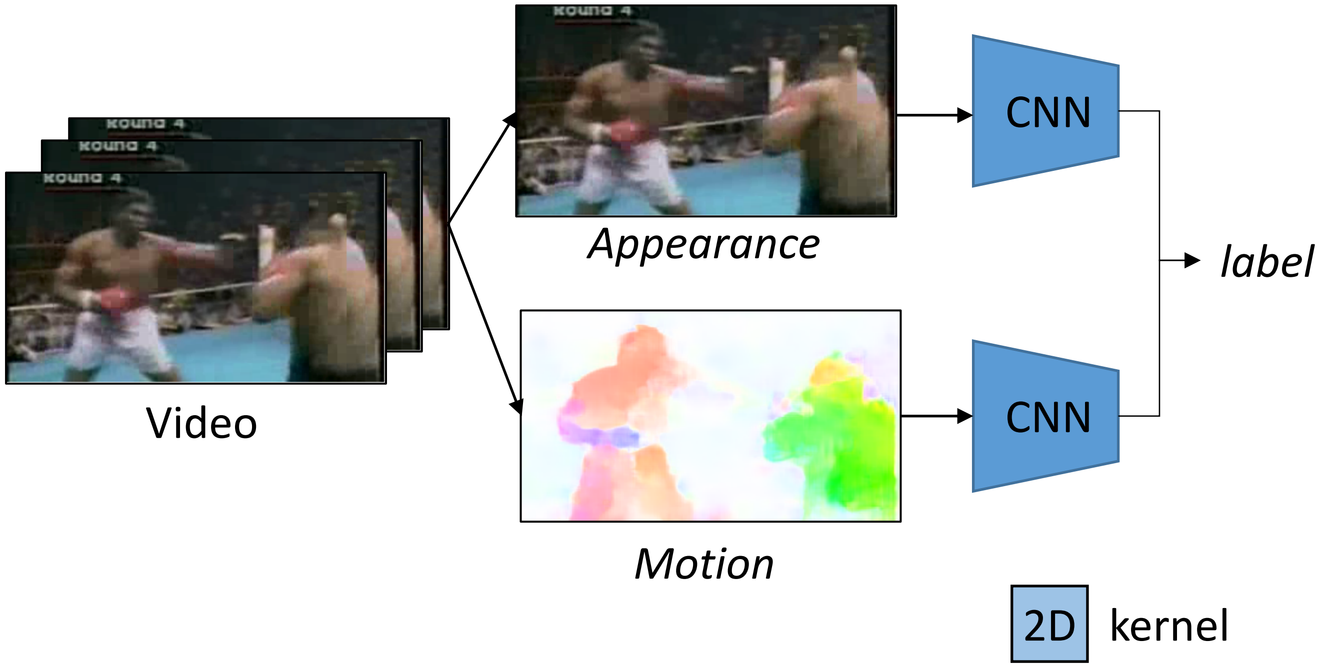
Context & Appearance

Swimming?

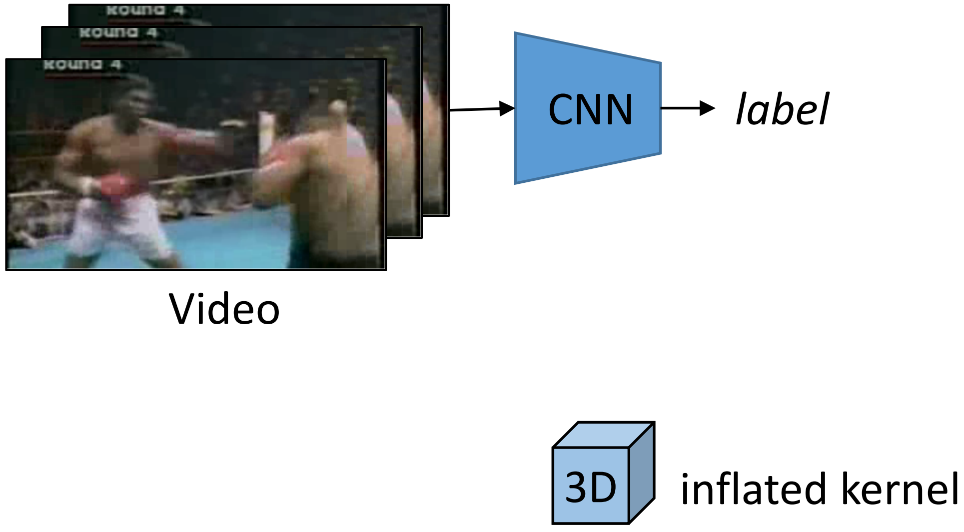
Handshaking

Action Recognition
Recent works

Two-stream



I3D



Limitations

Biased towards context

Lack of explainability

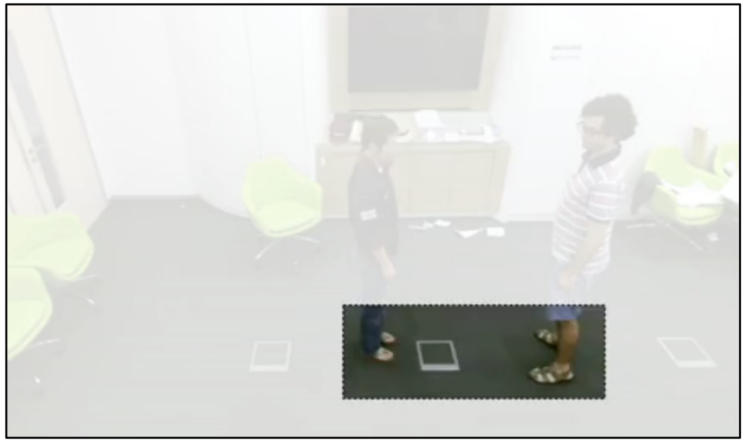
Human pose? Objects? Scene?

[Two-stream, Simonyan, NIPS'16]

[I3D, Carreira et al, CVPR'18]

Structured Deep Learning

Outline



Visual Attention

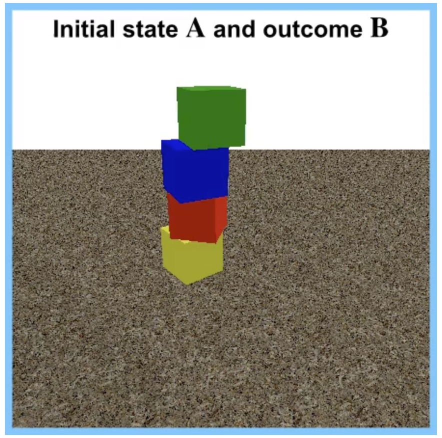
« Glimpse Clou
adel, C. Wol
G. Taylor
CVPR'18

Christian Wolf
INSA Lyon - LIRIS

Julien Mille
INSA CVL - LI Tours

Entity-level interactions

« Object level Reasoning »
F. Baradel, N. Neverova, C. Wolf, J. Mille, G. Mori
ECCV'18



Reasoning

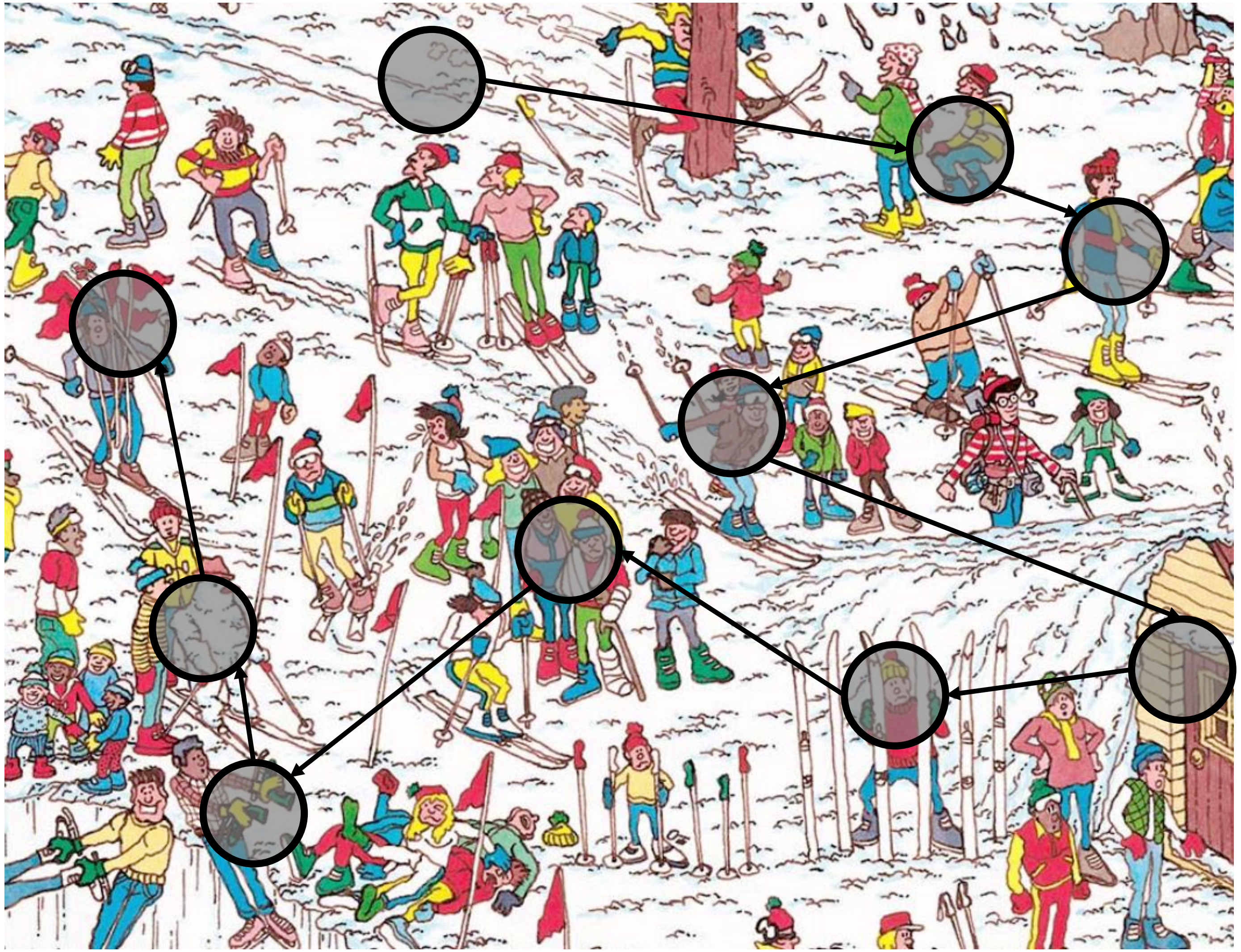
« Co factual learning »
F. Bar Neverova, J. Mille, ori, C. Wolf
ICLR'20 (spotlight)

Graham W. Taylor
University of Guelph
Vector Institute

Visual Attention

What is happening?

Winter activities

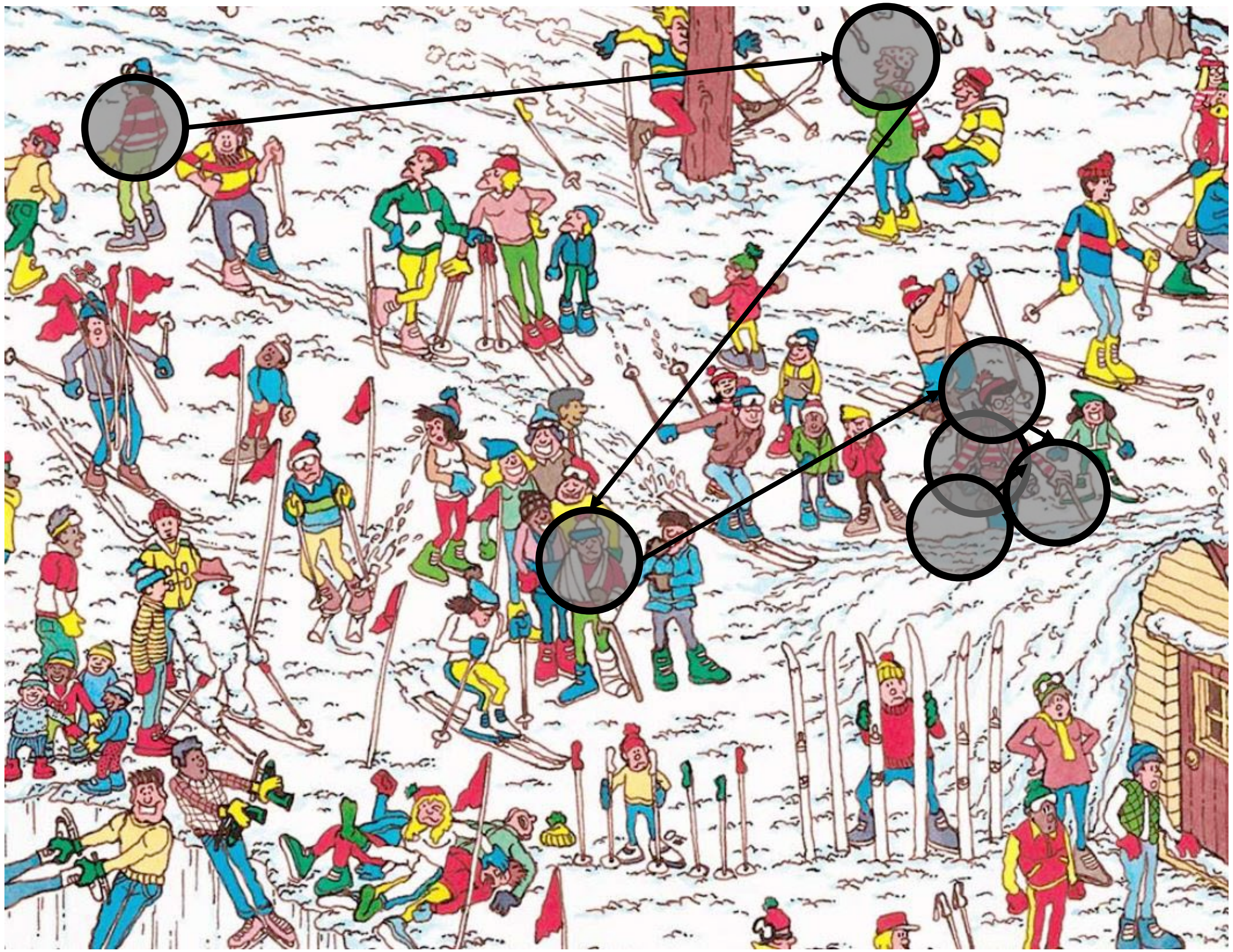


[Yarbus, 1976]
[Roger et al, 2012]

Visual Attention

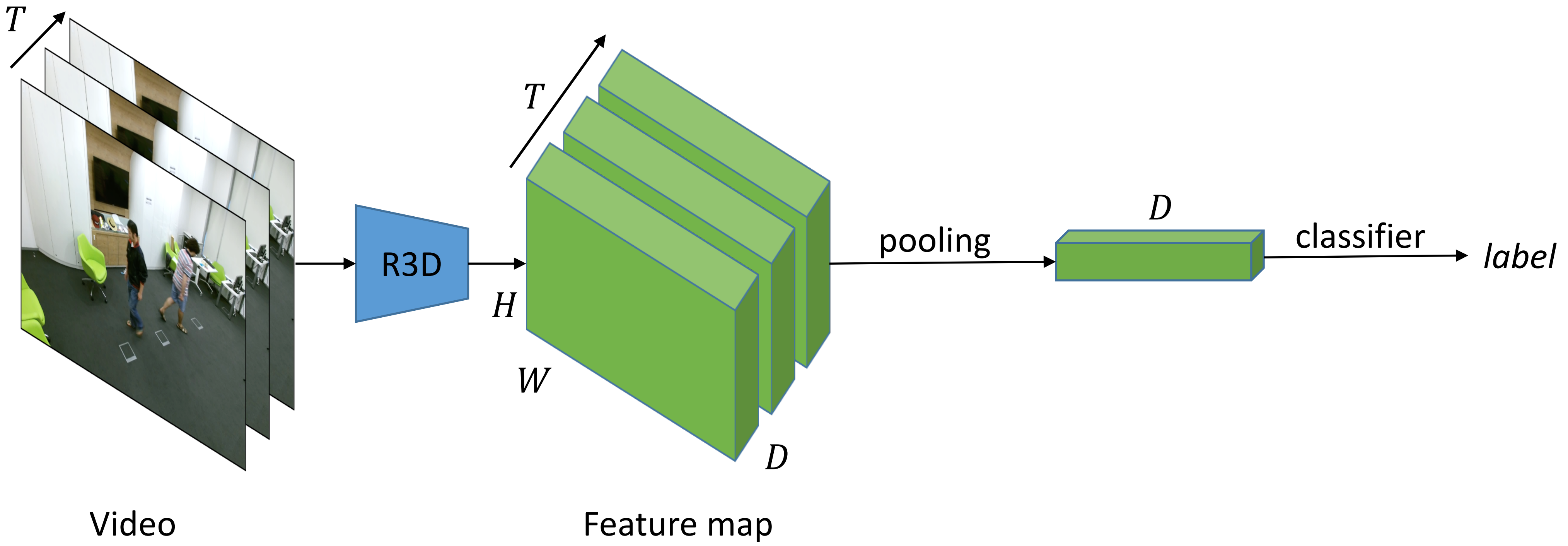
What is Charlie doing?

Walking



[Yarbus, 1976]
[Roger et al, 2012]

Action Recognition
Baseline

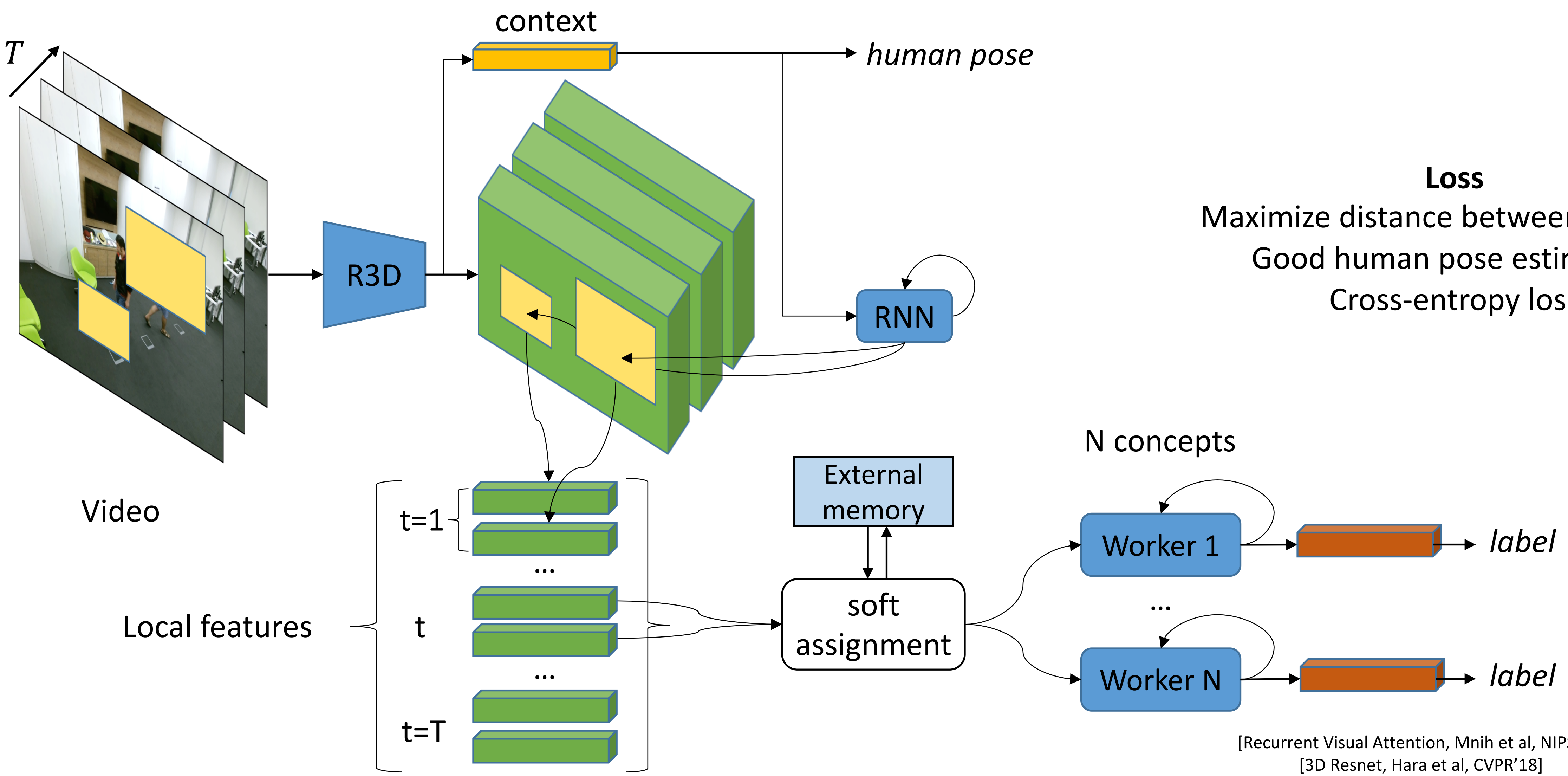


Limitations

What about fine-grained human actions?

How to focus on relevant parts of the video?

Glimpse Clouds
Method



Loss

Maximize distance between glimpses

Good human pose estimation

Cross-entropy loss

[Recurrent Visual Attention, Mnih et al, NIPS'14]

[3D Resnet, Hara et al, CVPR'18]

Glimpse Clouds
State-of-the-art results

Method	*Modality*	*CS*	*CV*
Ensemble TS-LSTM	skeleton	74.6	81.3
View invariant	skeleton	80.0	87.2
Hands-Attention (ours)	skeleton+ RGB	84.8	90.6
Glimpse Clouds (ours)	**RGB**	**88.4**	**93.2**

Accuracy on NTU-RGB+D

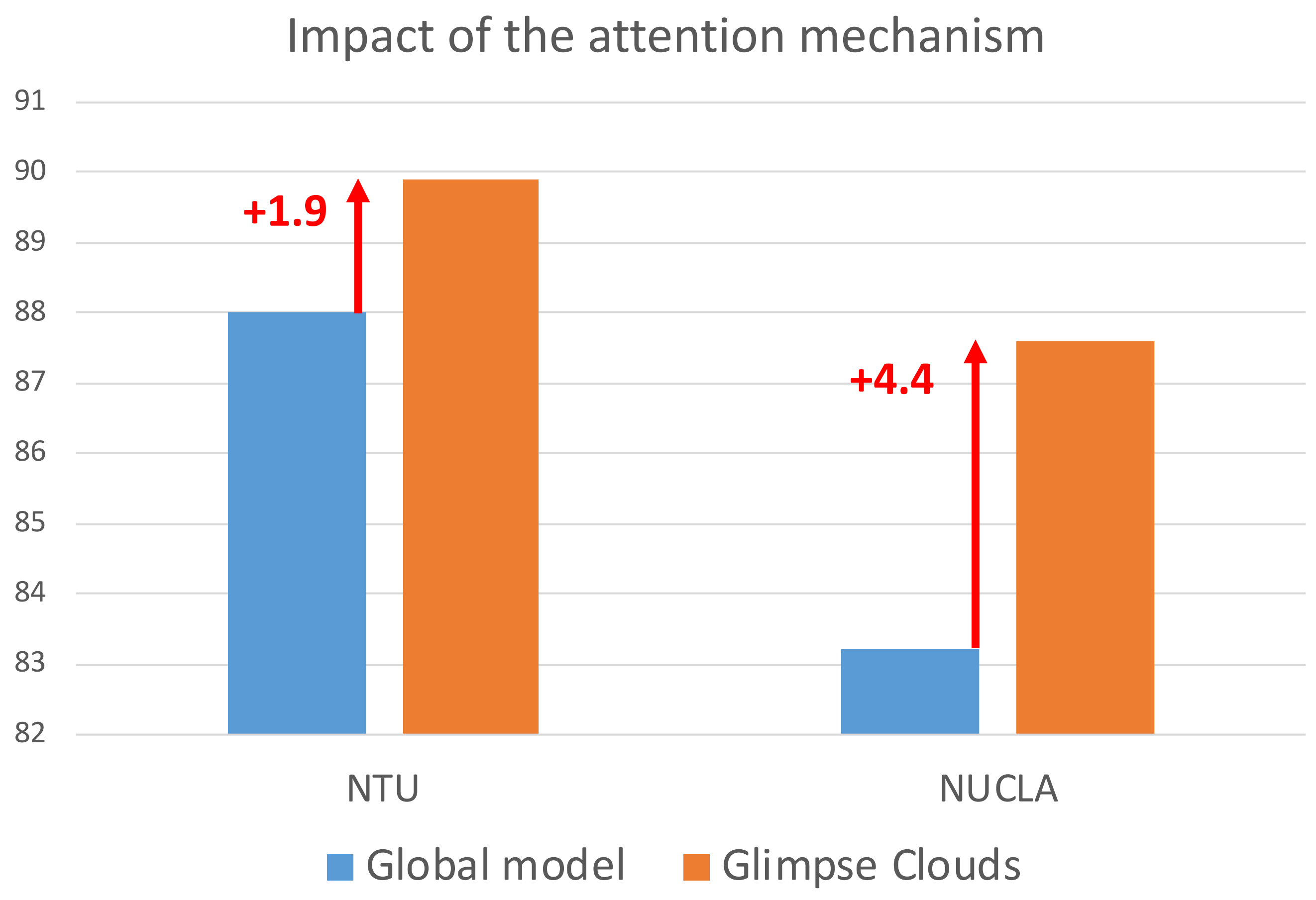
Method	*Modality*	*V1*
Enhanced viz.	Skeleton	86.1
Ensemble TS-LSTM	Skeleton	89.2
NKTM	RGB	75.8
Glimpse Clouds (ours)	**RGB**	**90.1**

Accuracy on Northwestern-UCLA

fabienbaradel/glimpse_clouds

[Pose-driven hands-attention, Baradel et al, BMVC'18]

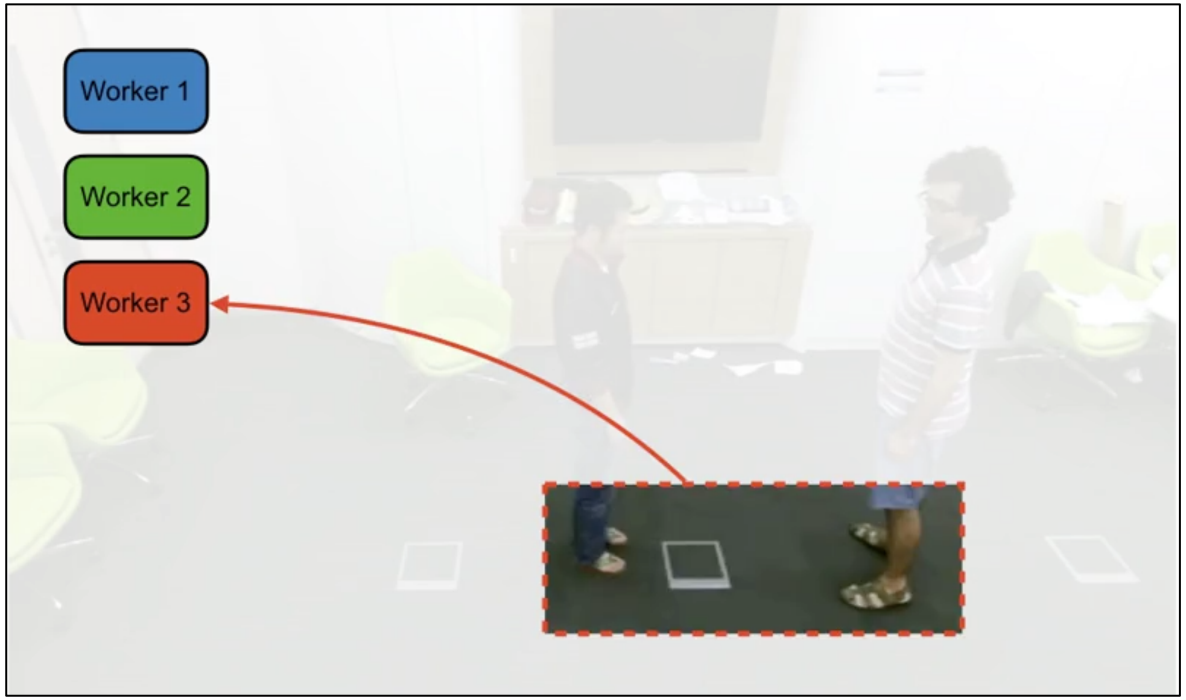
Glimpse Clouds
Ablation study



Resolution matters
Local fine-grained features

Glimpse Clouds
Visualization

Raw video

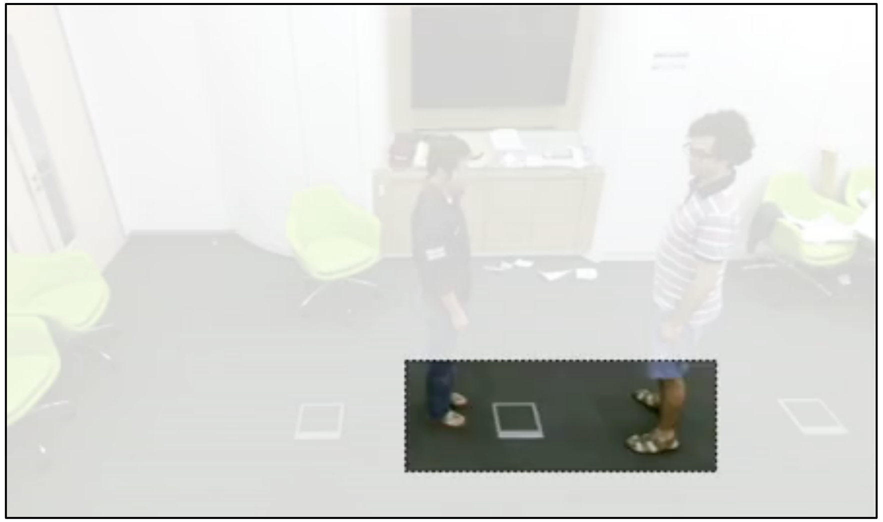


Attended regions

Worker 1 → ~Hands
Worker 2 → ~Heads
Worker 3 → ~Legs

Note: argmax shown for the feature-to-worker association

Outline



Unstructured local features…
Incorporate structure from images?
Leverage visual entities interactions?

Visual Attention

Christian Wolf
INSA Lyon - LIRIS

Julien Mille
INSA CVL - LI Tours

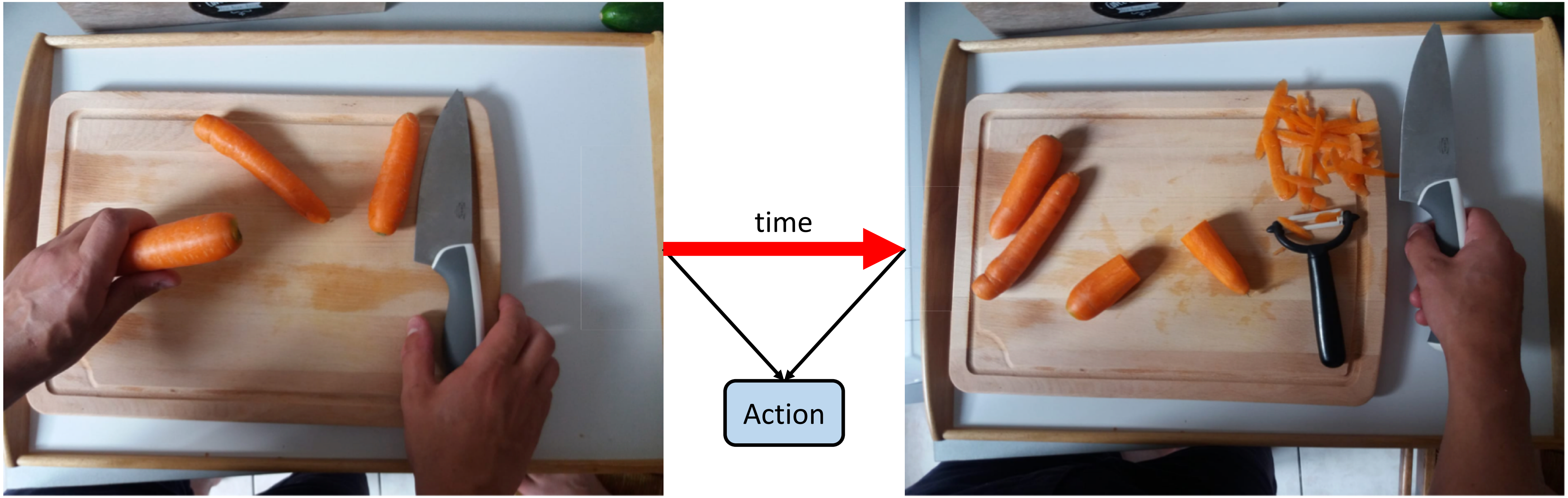
Entity-level interactions

« …vel Re…
F. Ba… Nevero… …olf,
J. Mille, G. Mori
ECCV'18

Natalia Neverova
Facebook

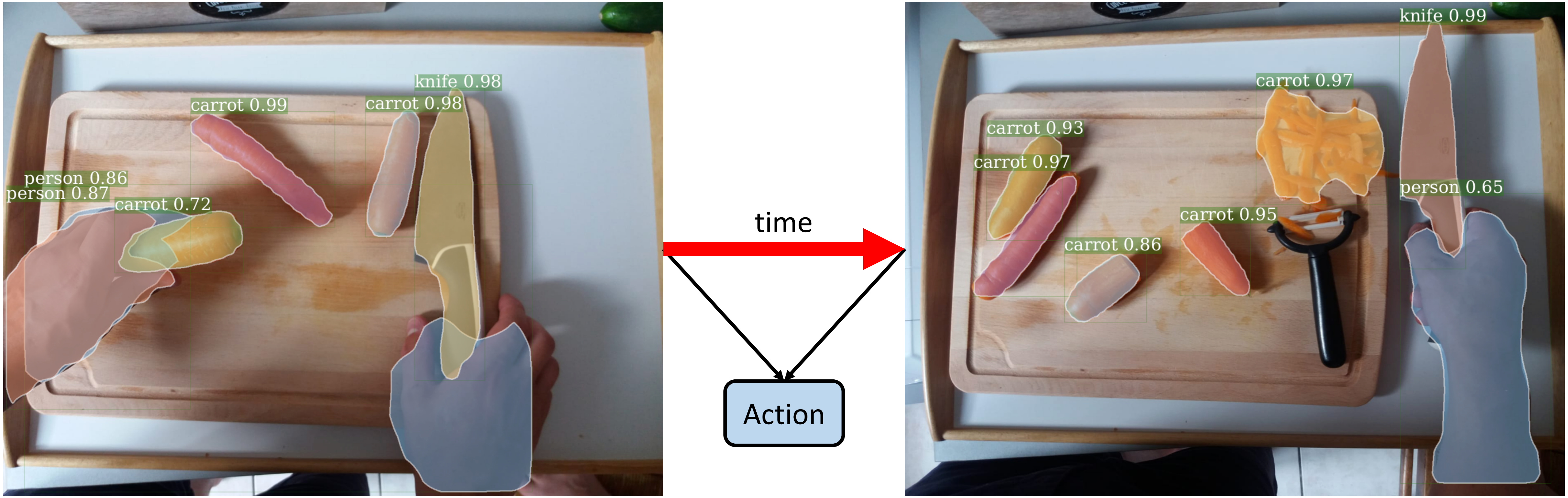
Greg Mori
SFU

Object-level Reasoning



Often possible to infer what happened from few frames

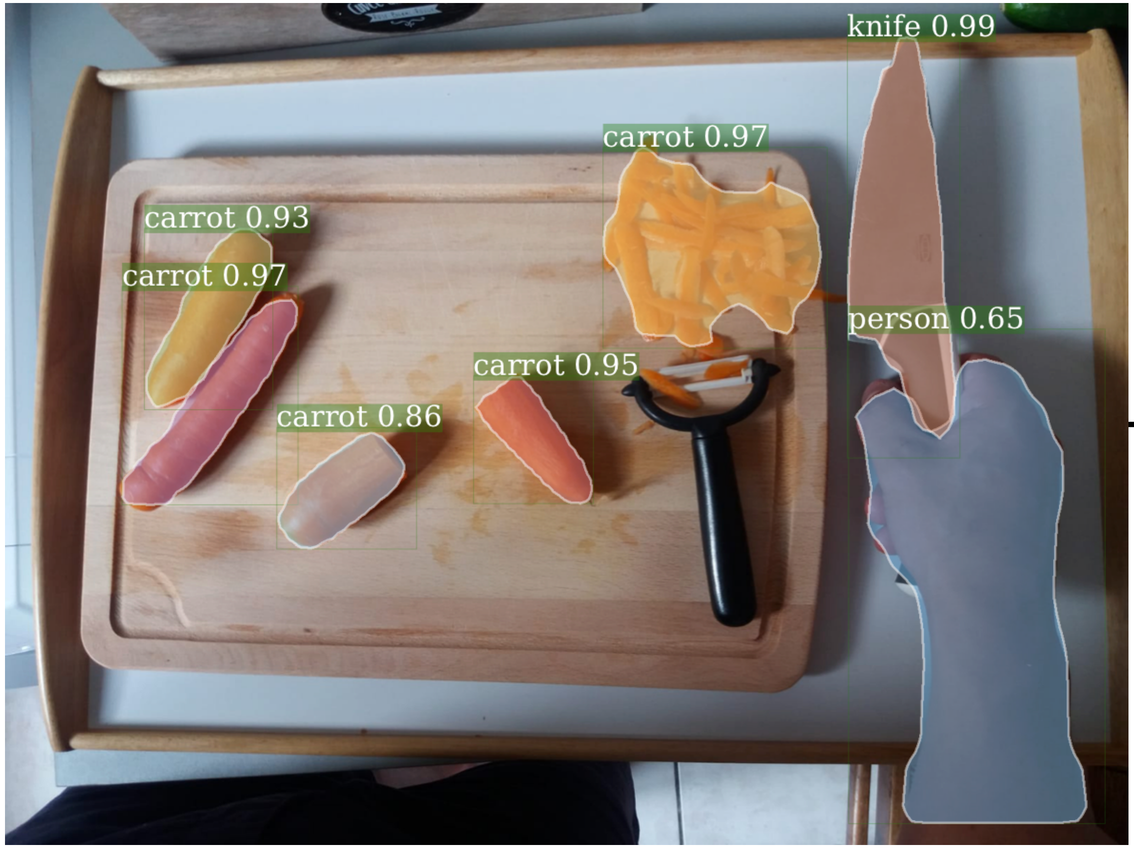
Object-level Reasoning



Often possible to infer what happened from few frames
Visual entities interactions

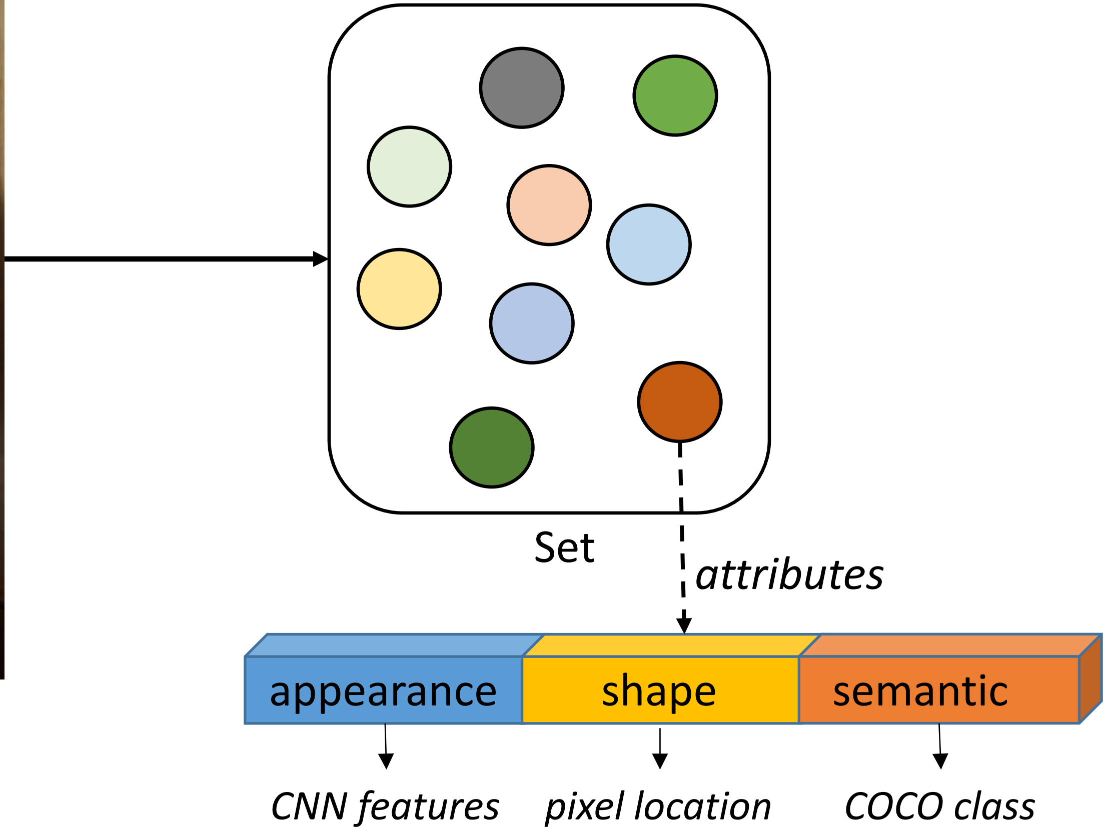
[Mask-RCNN, He et al, ICCV’17]

Image as a set of objects

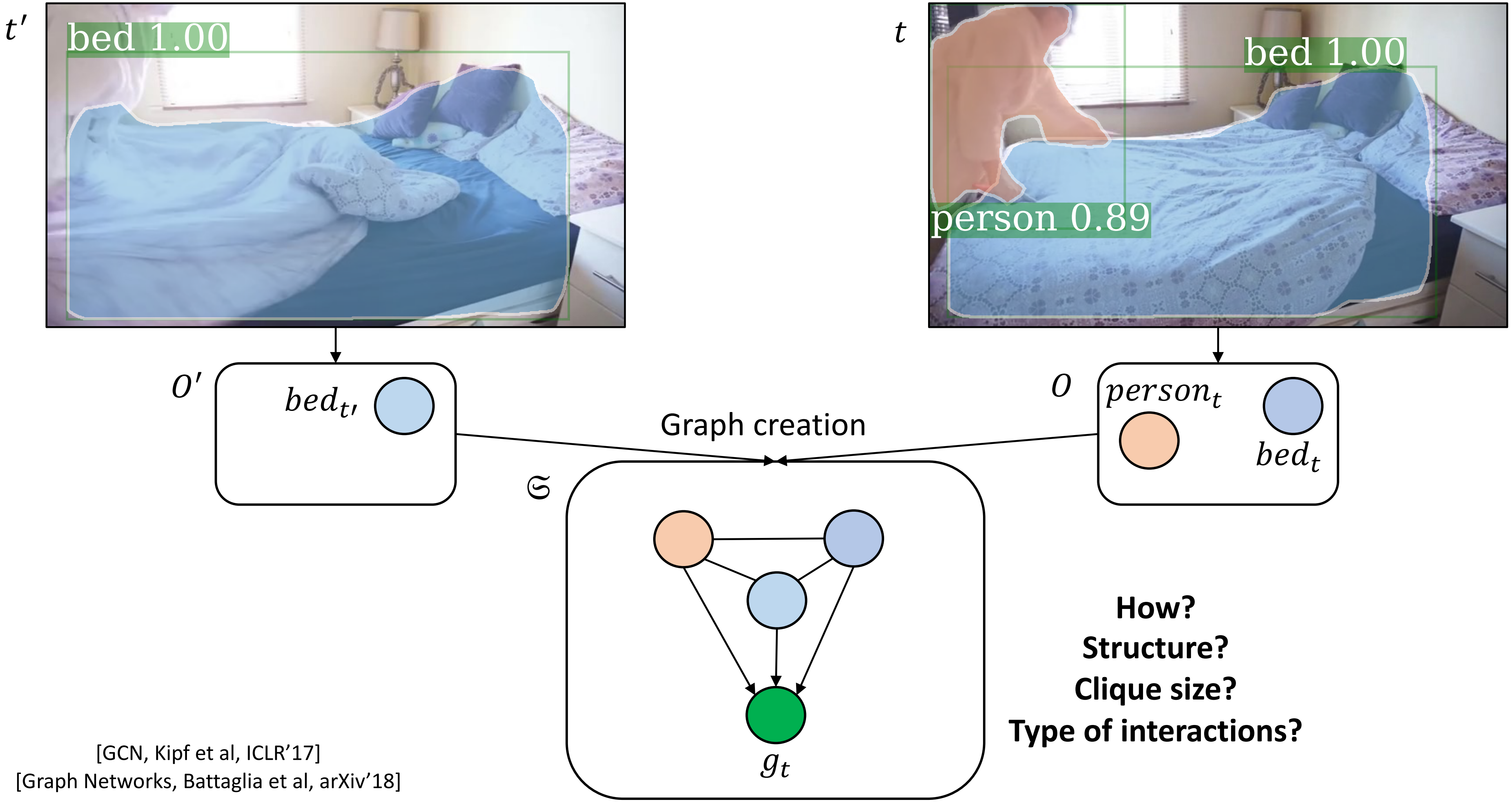


RGB

Mask-RCNN



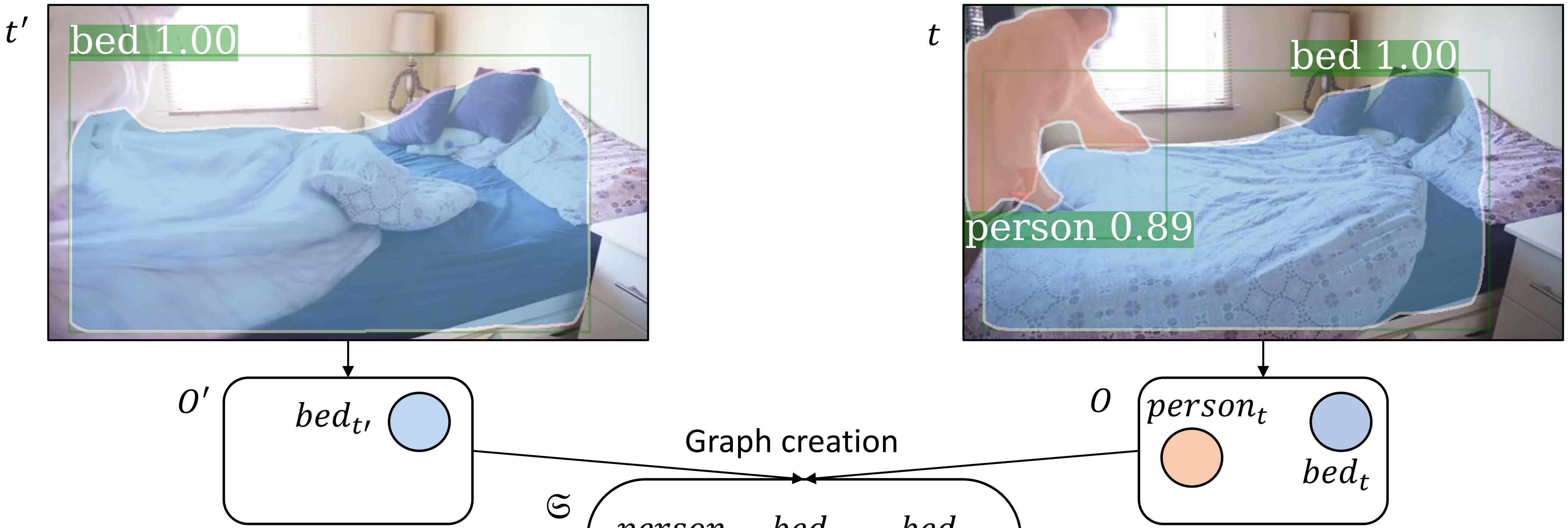
Object Relation Network

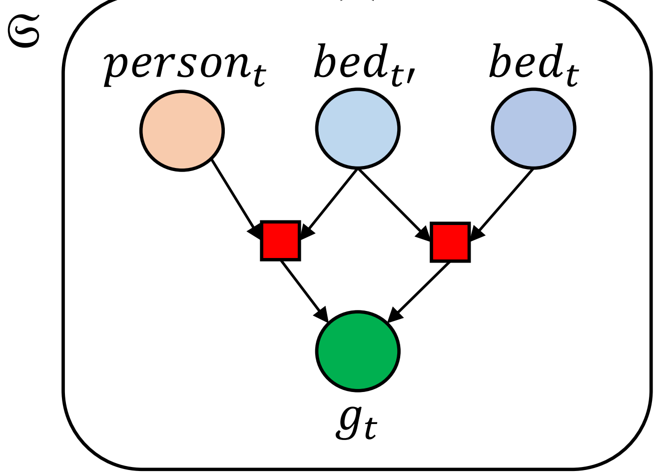


How?
Structure?
Clique size?
Type of interactions?

[GCN, Kipf et al, ICLR'17]
[Graph Networks, Battaglia et al, arXiv'18]

Object Relation Network





Data efficient
Invariant to the number of objects
Inter-frame object relations
Semantic meaning
Clique of size 2

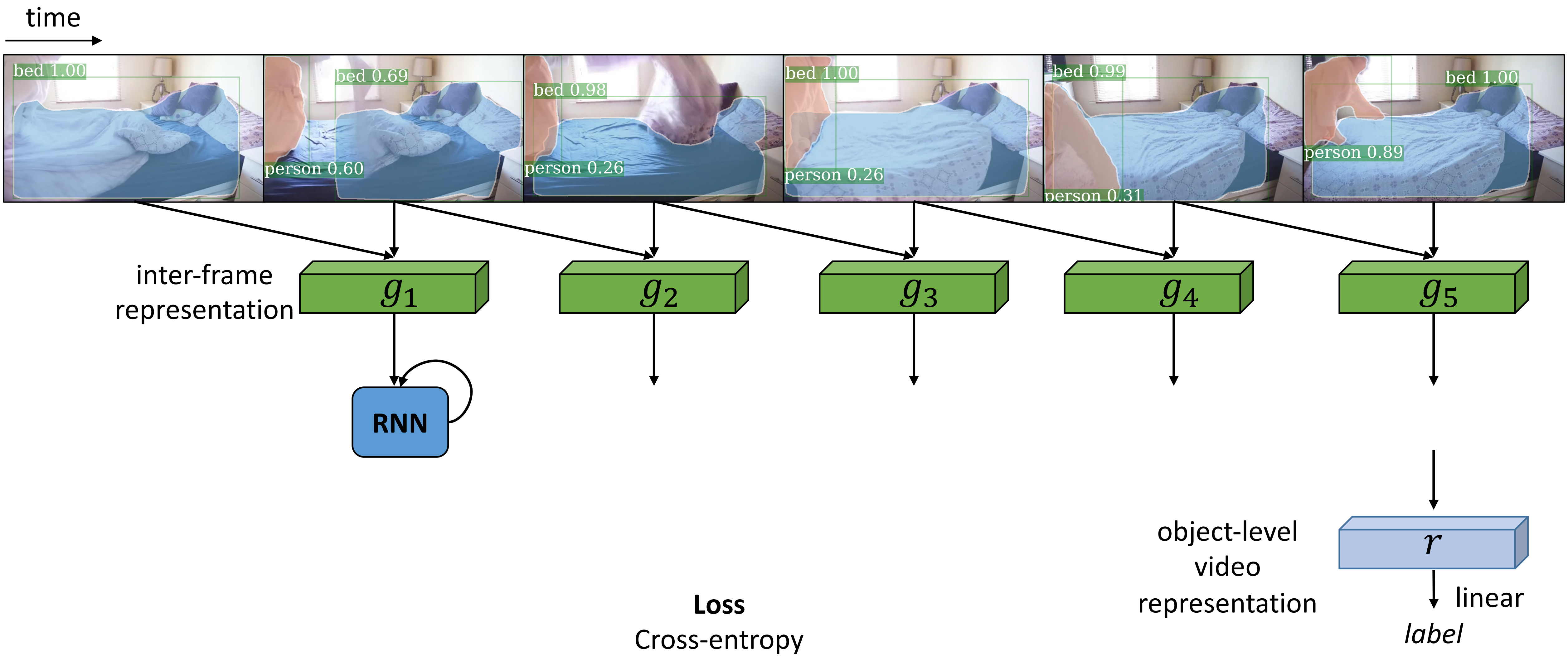
Shared MLP

$$g_t = f(bed_{t'}, person_t) + f(bed_{t'}, bed_t)$$

$$g_t = \sum_{o \in O} \sum_{o' \in O'} f(o', o)$$

[GCN, Kipf et al, ICLR'17]
[Graph Networks, Battaglia et al, arXiv'18]

Object Relation Network



Object Relation Network
State-of-the-art

Method	*Acc.*
C3D	21.50
I3D	27.63
Multiscale TRN	33.60
Object Relation Network	**35.97**

Accuracy on Something-Something

Method	*mAP*
Resnet50	40.5
I3D	39.7
Object Relation Network	**44.7**

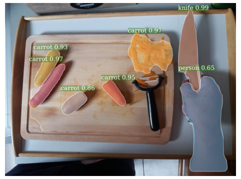
Mean Average Precision on VLOG

Method	*Acc.*
Resnet18	32.05
Resnet3D-18	34.20
Object Relation Network	**40.89**

Verb accuracy on EPIC Kitchens

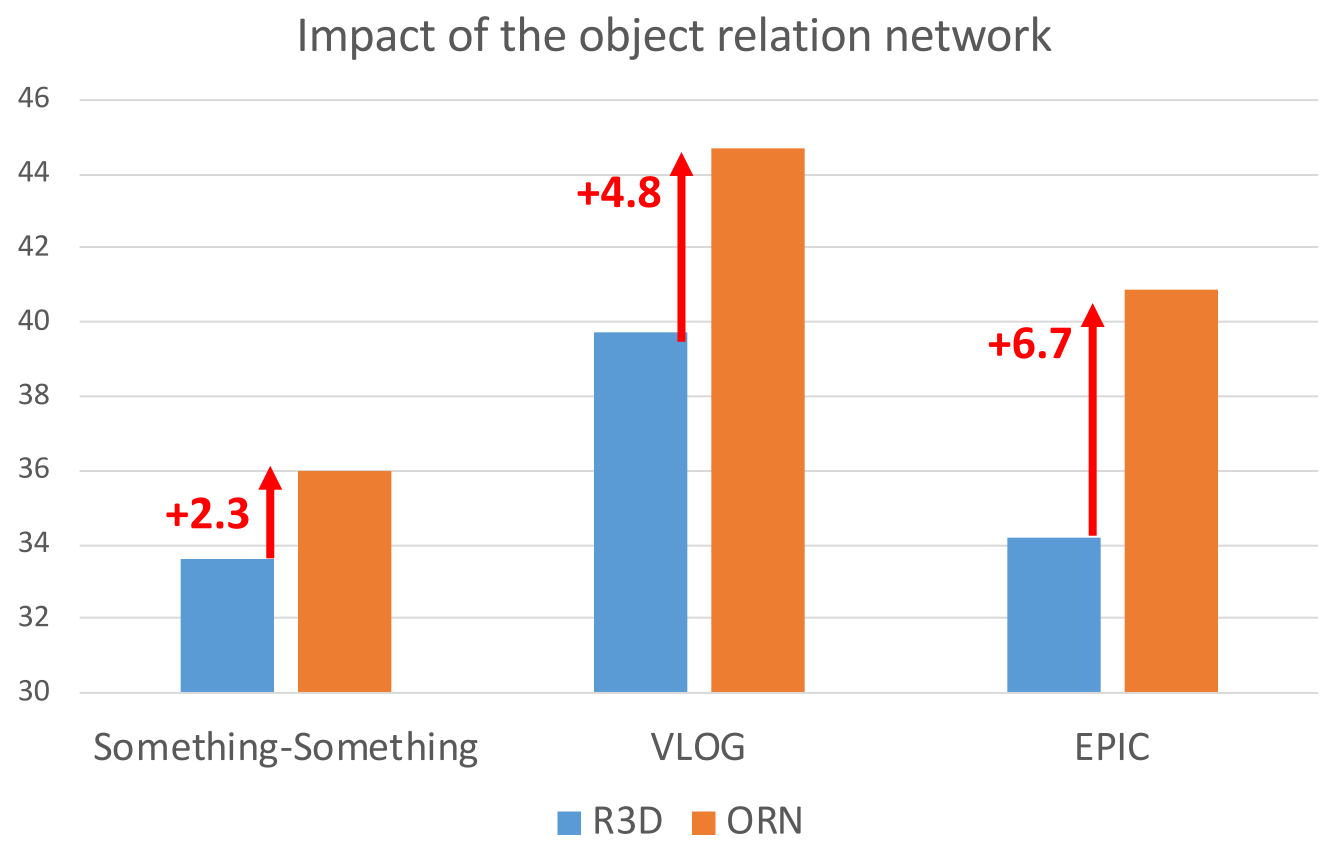
fabienbaradel/object_level_visual_reasoning

+



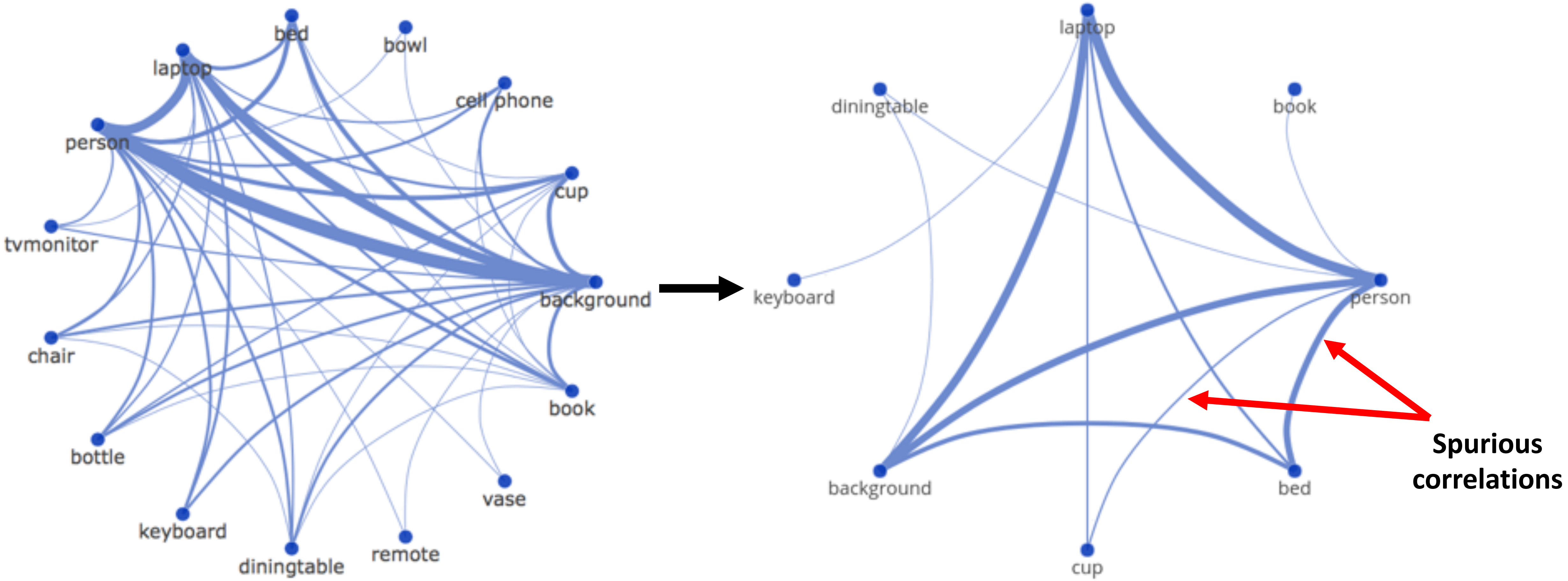
Object masks detected by Mask-RCNN

Object Relation Network
Ablation study



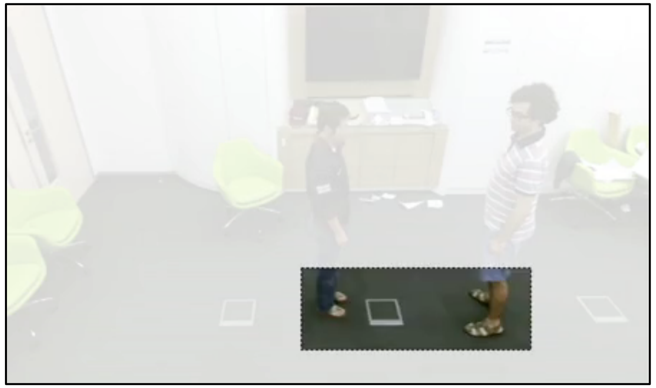
Detection performance
High resolution

Object Relation Network
Interactions



Co-occurences *Human-Laptop interactions* Learned relations

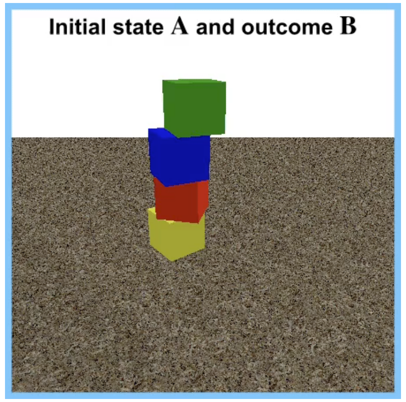
Outline



Visual Attention

Entity-level interactions

Structure matters…
Can we go one step further?
Beyond supervised learning?
Learning underlying latent concepts?



Reasoning

« Counterfactual learning »
F. Baradel, N. Neverova, J. Mille, C. Wolf
ICL light)

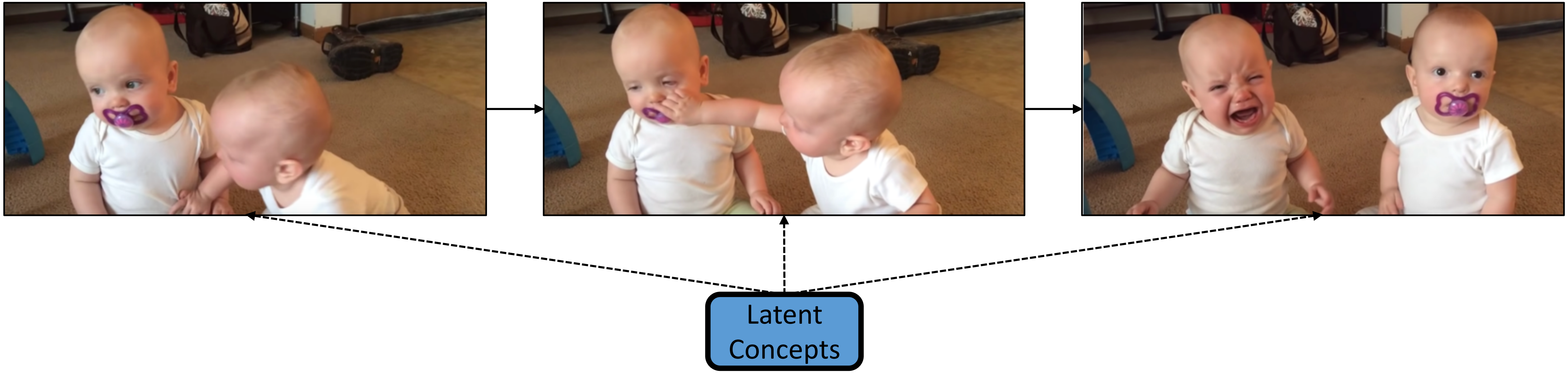
Christian Wolf
INSA Lyon - LIRIS

Julien Mille
INSA CVL - LI Tours

Natalia Neverova
Facebook

Greg Mori
SFU

Reasoning & Causation



Understanding of complex relationships
Cause-effect

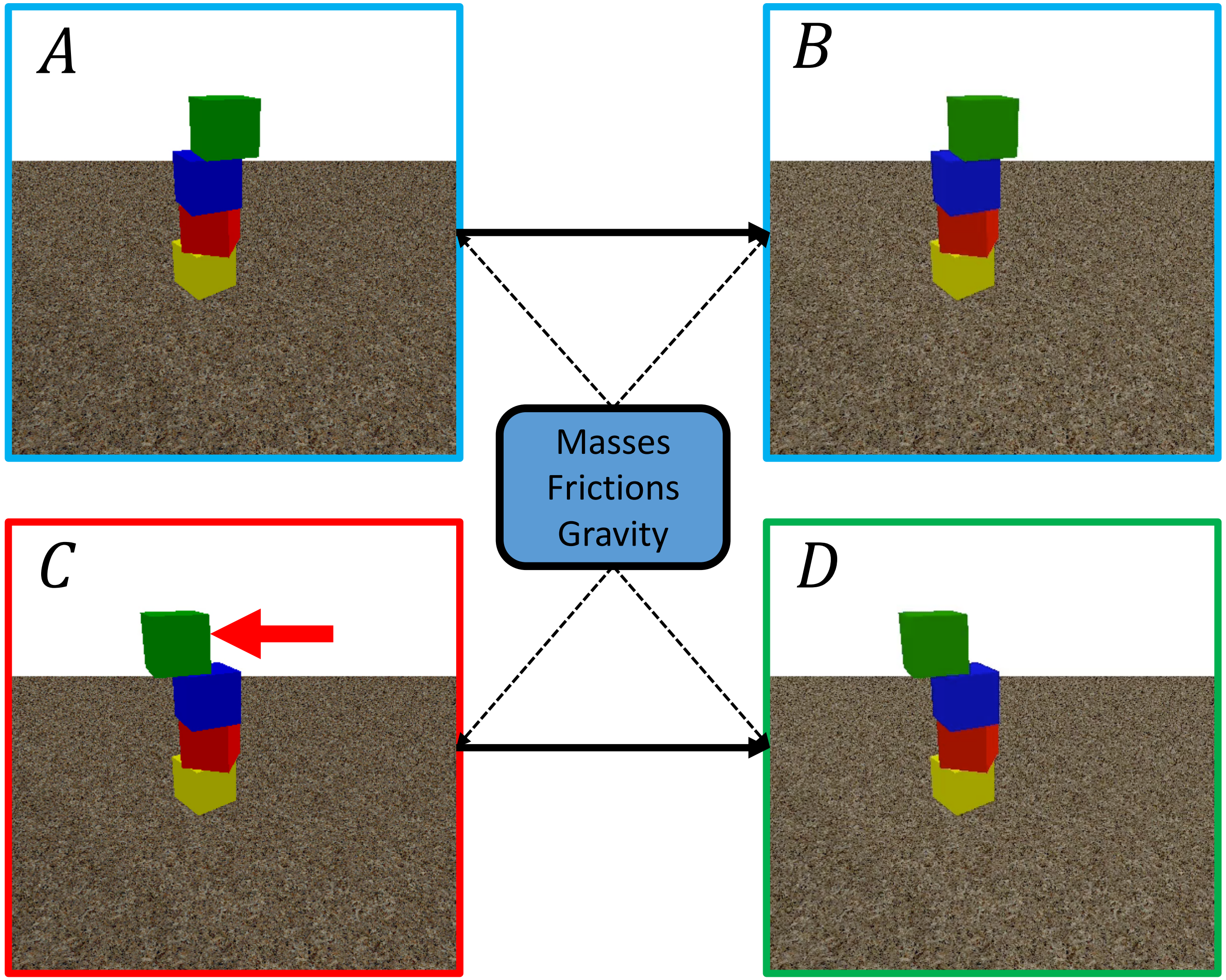
What would happened if?
Counterfactual statement

Counterfactual

Future forecasting

Initial state

Outcome



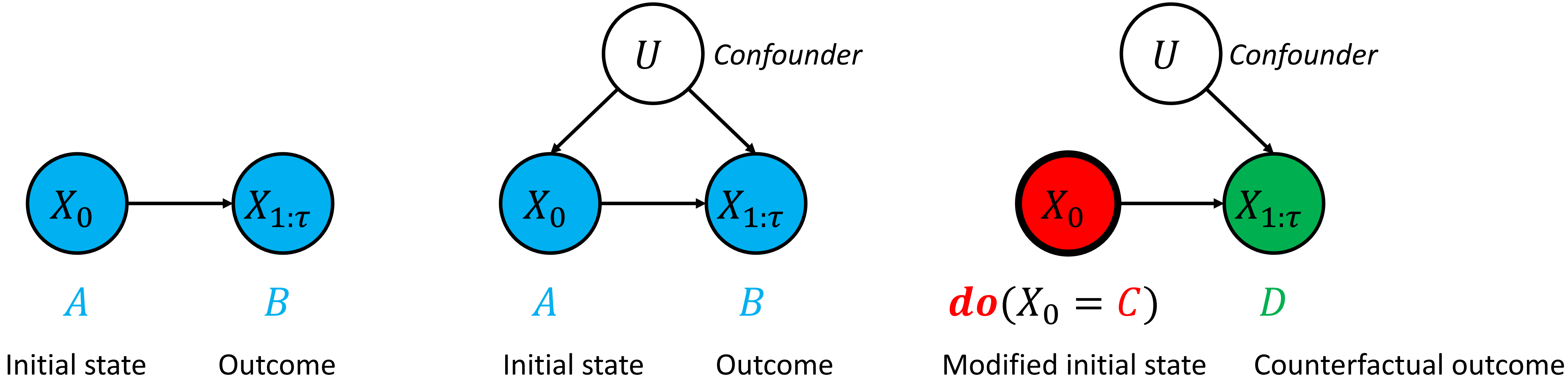
Modified initial state

Counterfactual outcome

Counterfactual

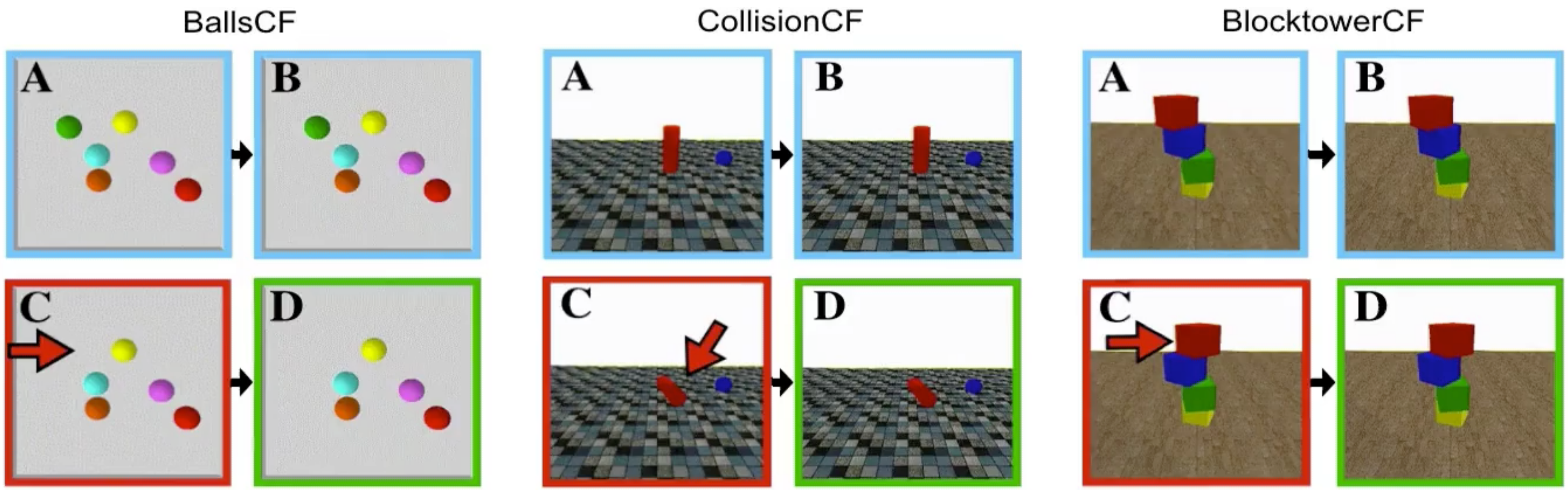
Future forecasting

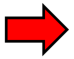
Feedforward **Counterfactual**

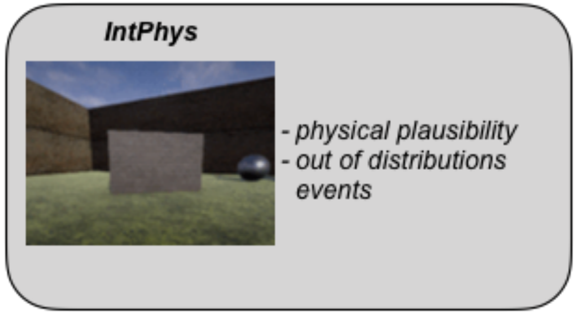


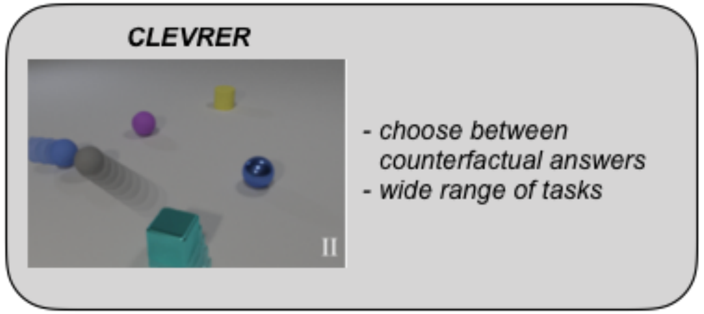
[Algorithmization of counterfactuals, Pearl, arXiv'18]

CoPhy benchmark



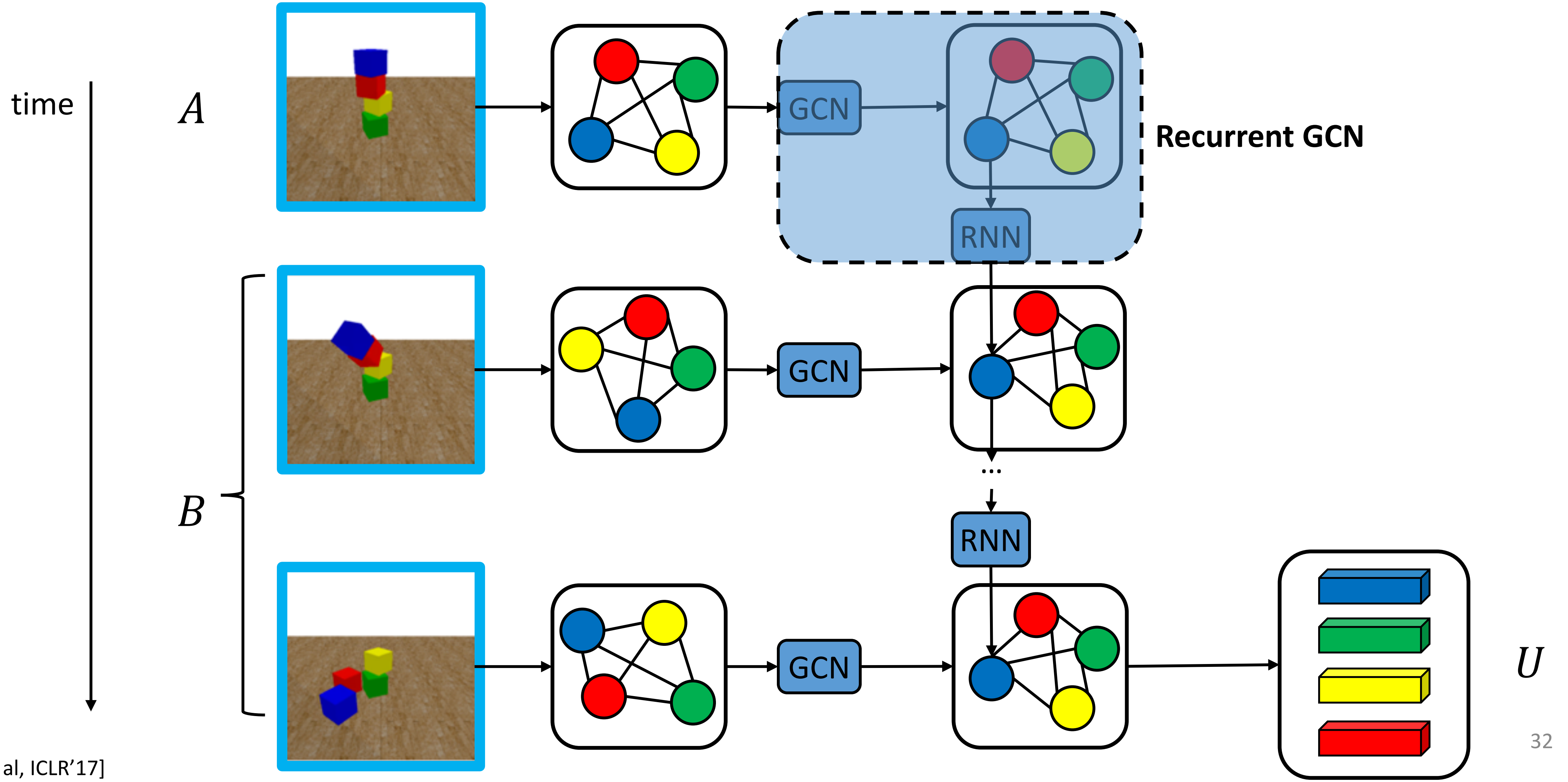
Large-scale datasets
250k examples ((A,B), (C,D))
7 millions of frames
Supervision of the do-operator ()
Confounders are necessary for future prediction





CoPhyNet

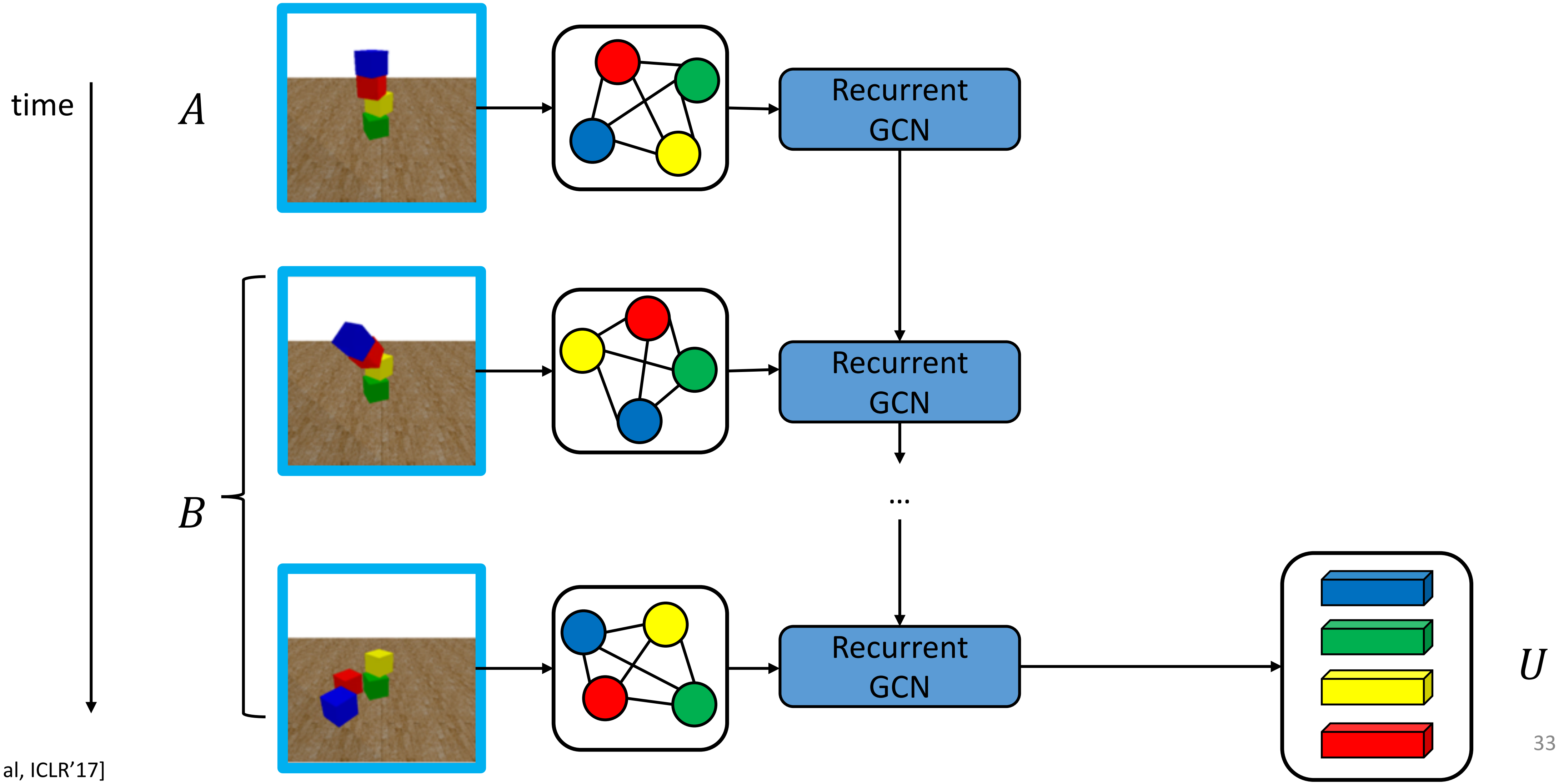
Unsupervised confounders estimations



[GCN, Kipf et al, ICLR'17]

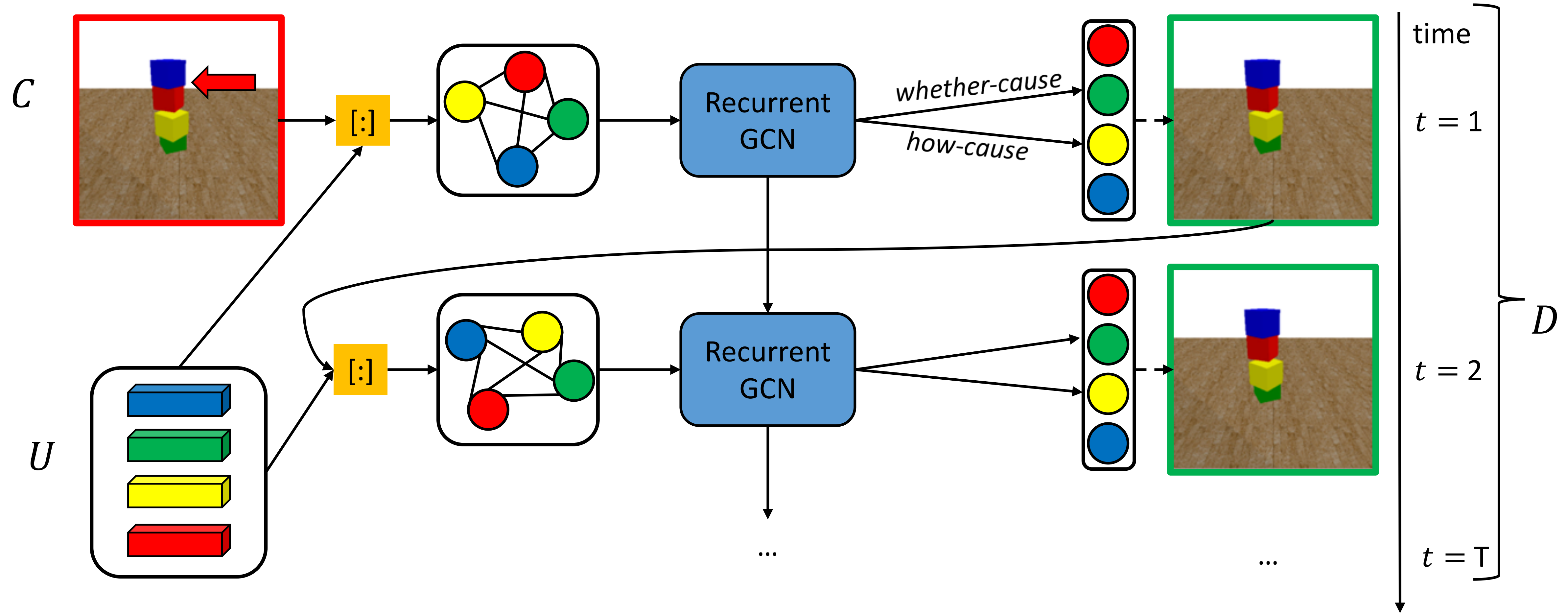
CoPhyNet

Unsupervised confounders estimations



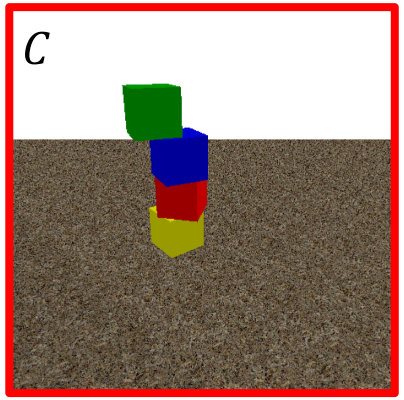
[GCN, Kipf et al, ICLR’17]

CoPhyNet
Trajectory prediction



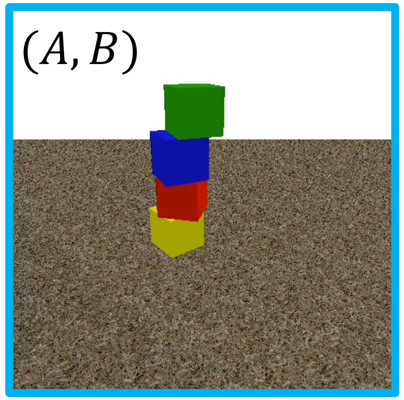
[Perceived-causality, Gerstenberg et al, ACCSS'15]

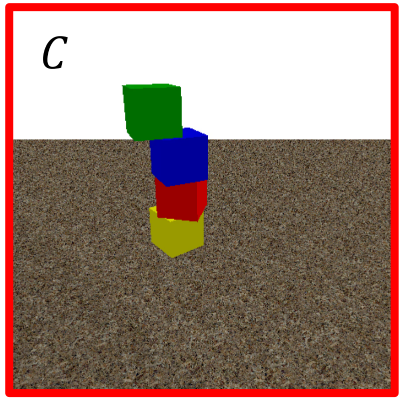
Human study



?

Human non-CF

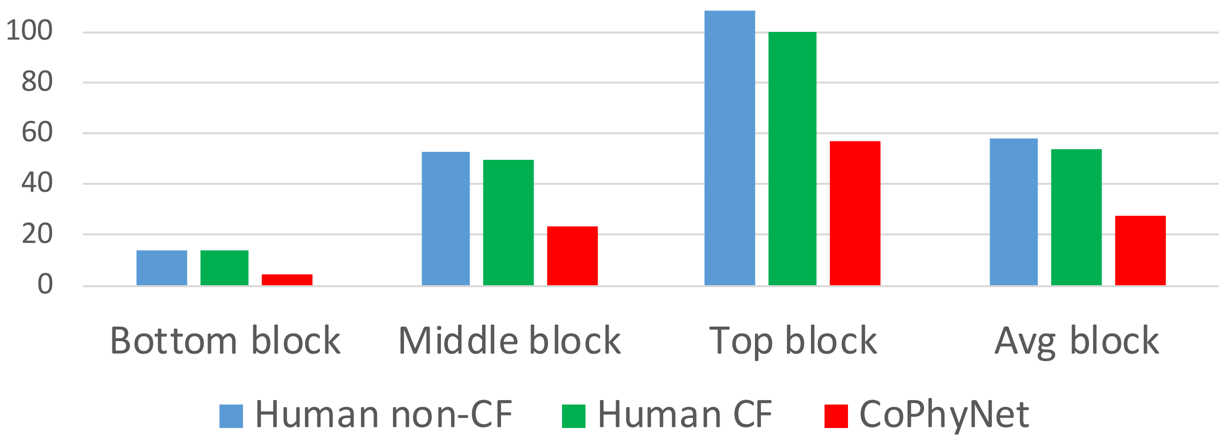




?

Human CF

2D pixel error for each block



Cophynet

Results

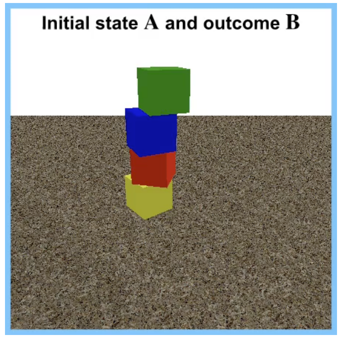
NOT COMPARABLE!

	Train → Test	Copying baselines: Copy C	Copying baselines: Copy B	Feedforward models: IN	Feedforward models: NPE	CoPhyNet	Soft-upper bound: IN Sup.
	3 → 3	0.470	0.601	0.318	0.331	**0.294**	*0.296*
Unseen confounders	3 → 3*	0.365	0.592	0.298	0.319	**0.289**	*0.282*
Unseen number of blocks	3 → 4	0.754	0.846	0.524	0.523	**0.482**	*0.467*

MSE on 3D positions (average over time)

fabienbaradel/cophy

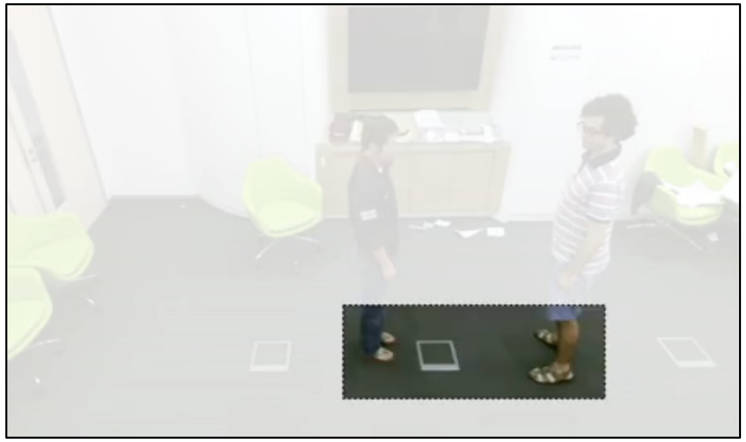
+

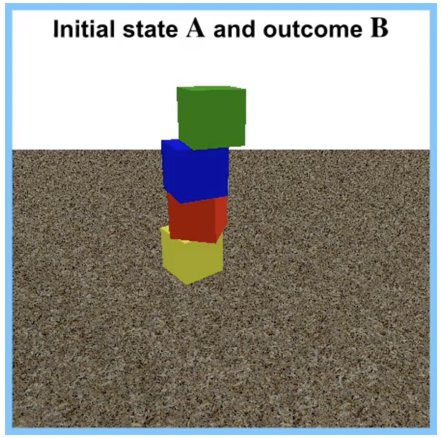


CoPhy benchmark

[Interaction Network, Battaglia et al, NIPS'17]
[Neural Physic Engine, Chang et al, ICLR'17]

Conclusion





Visual Attention

« Glimpse Clouds »
F. Baradel, C. Wolf, J. Mille, G. Taylor, **CVPR'18**

Focus on important parts
Automatic selection
Distributed recognition

Entity-level interactions

« Object level Reasoning »
F. Baradel, N. Neverova, C. Wolf, J. Mille, G. Mori, **ECCV'18**

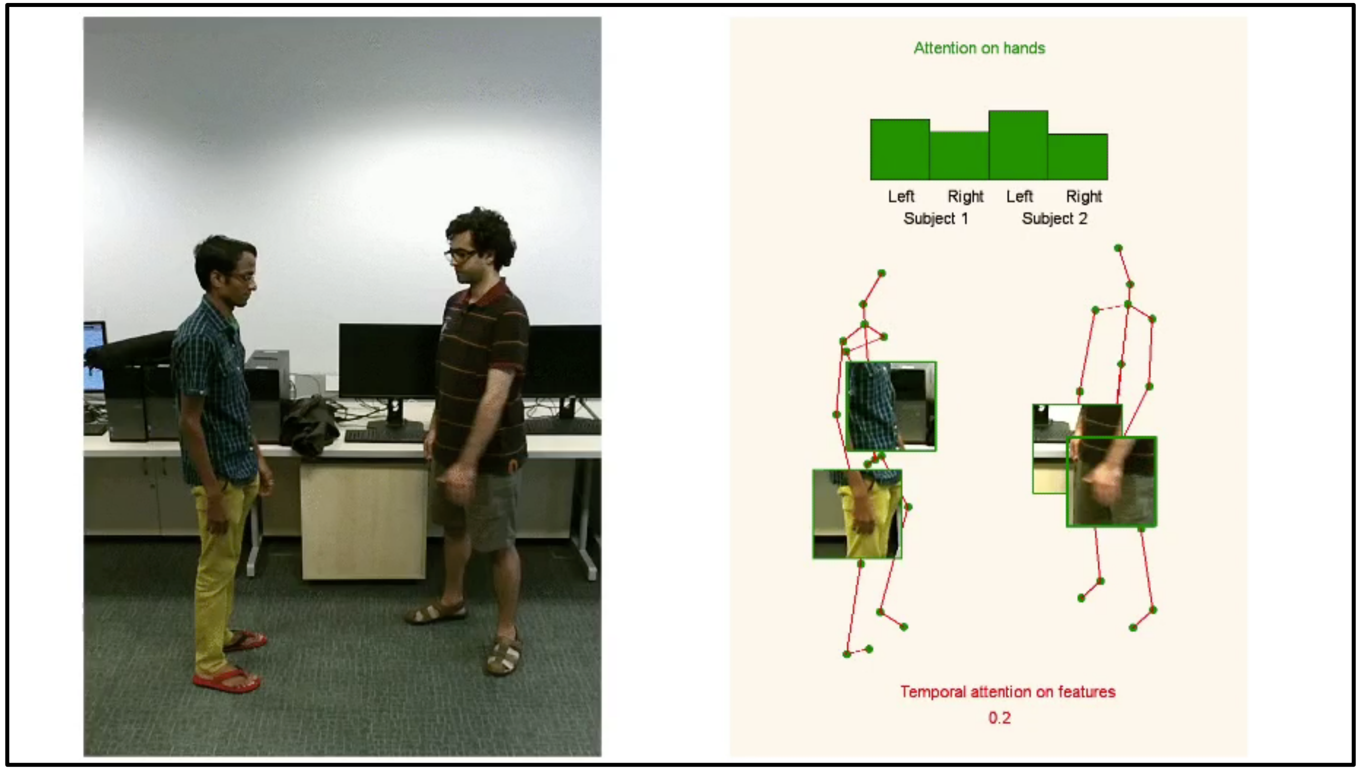
Object-centric modeling
Intra-time interactions
Learned relations

Reasoning

« Counterfactual learning »
F. Baradel, N. Neverova, J. Mille, G. Mori, C. Wolf, **ICLR'20 (spotlight)**

Unsupervised latent discovery
Future trajectory
New task in visual space

Other works
Pose-driven Attention

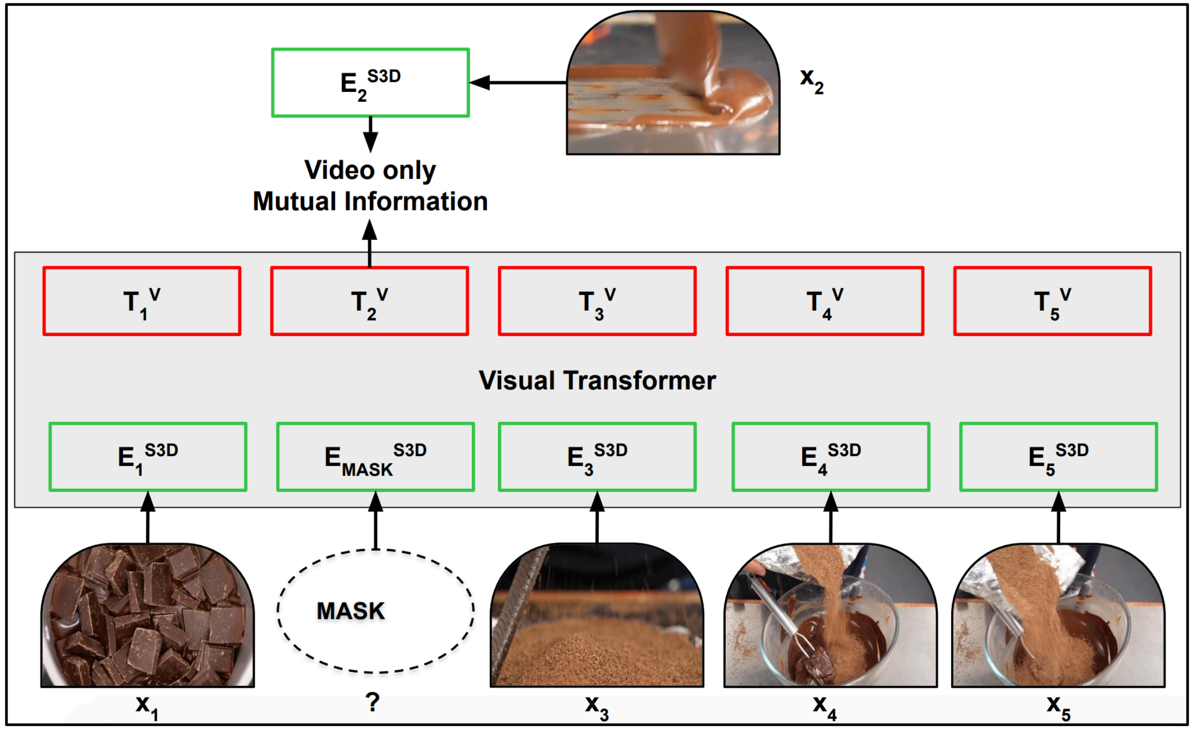


« Focus to Hands»
F. Baradel, C. Wolf, J. Mille,
ICCVW'17

Skeleton & RGB
Handcrafted/Learned context features
Focus around hands

« Pose-driven Attention to RGB»
F. Baradel, C. Wolf,J. Mille,
BMVC'18

Other works
Self-supervised learning



« Contrastive Bidirectionnal Transformer »
C. Sun, F. Baradel, K. Murphy, C. Schmid
Under review ECCV'20

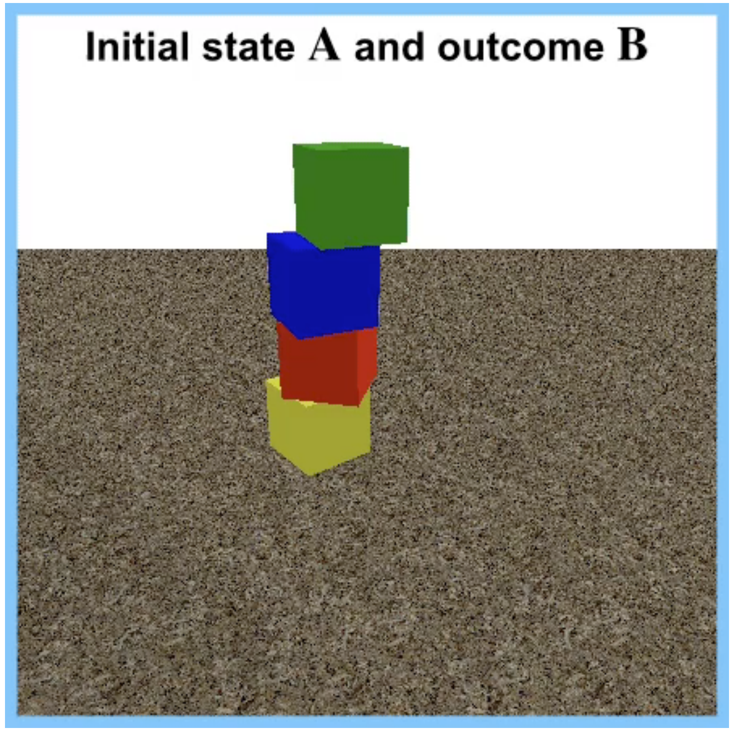
Human learning
Beyond large-scale annotated datasets
Efficient
Discover regularities
Prediction of missing parts
Instructional videos
Vision-Text alignment

Cordelia Schmid
INRIA – Google

Chen Sun
Google

Kevin P. Murphy
Google

What next?
Real word counterfactual predictions



« CoPhy++ »

F. Baradel, N. Neverova, J. Mille, G. Mori, C. Wolf

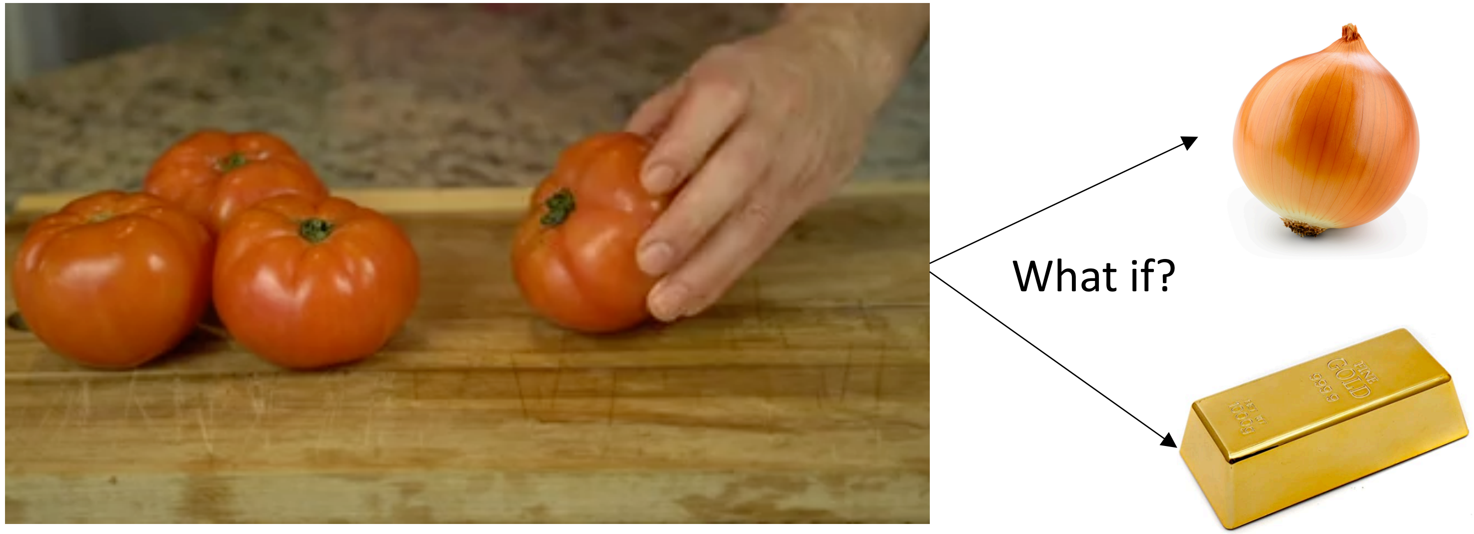
To be submitted to TPAMI

No object supervision

Unsupervised keypoints

Predictions in image space

What next?
Real word counterfactual predictions



Beyond correlation and dataset biases
Latent concept
Generalization
Semantical structure
Ontology

What next?
Disentangled representation

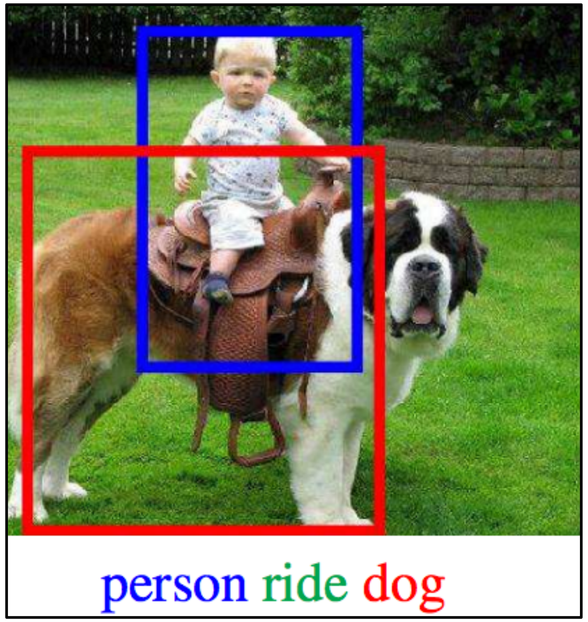


Image vs Video
Appearance vs Motion
Human Object Interaction vs Long-range activities
Efficient representation

Thank you!

Christian Wolf
INSA Lyon - LIRIS

Julien Mille
INSA CVL - LI Tours

Natalia Neverova
Facebook AI Research

Graham W. Taylor
University of Guelph
Vector Institute

Greg Mori
Simon Fraser University
Borealis AI

Cordelia Schmid
INRIA – Google

Chen Sun
Google

Kevin P. Murphy
Google