OBJECT LEVEL VISUAL REASONING IN VIDEOS

Fabien Baradel[1], Natalia Neverova[2], Christian Wolf[1,3], Julien Mille[4], Greg Mori[5]

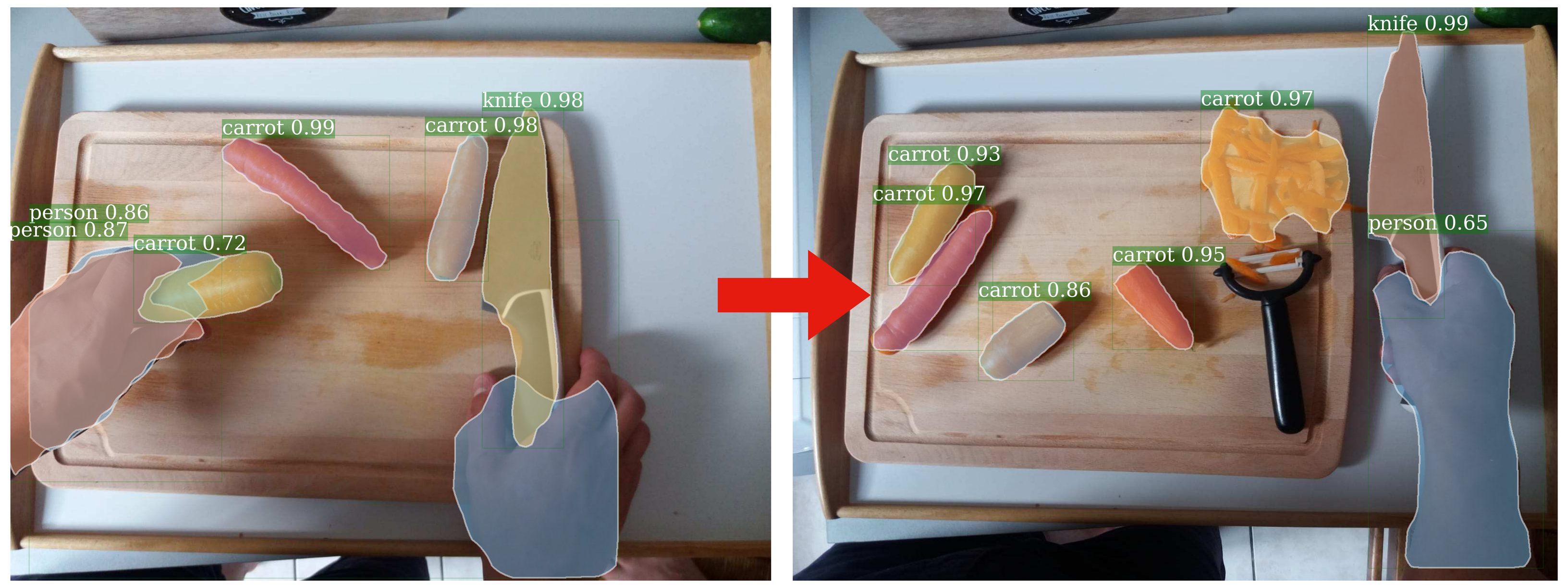
[1]INSA Lyon – LIRIS, [2] Facebook AI Research, [3] INRIA-Chroma, CITI lab, [4] LIFAT, [5] SFU

Contributions

- Reasoning over semantic structures
- Relations between detected objects
- Spatio-temporal object interactions
- State-of-the art on 3 datasets

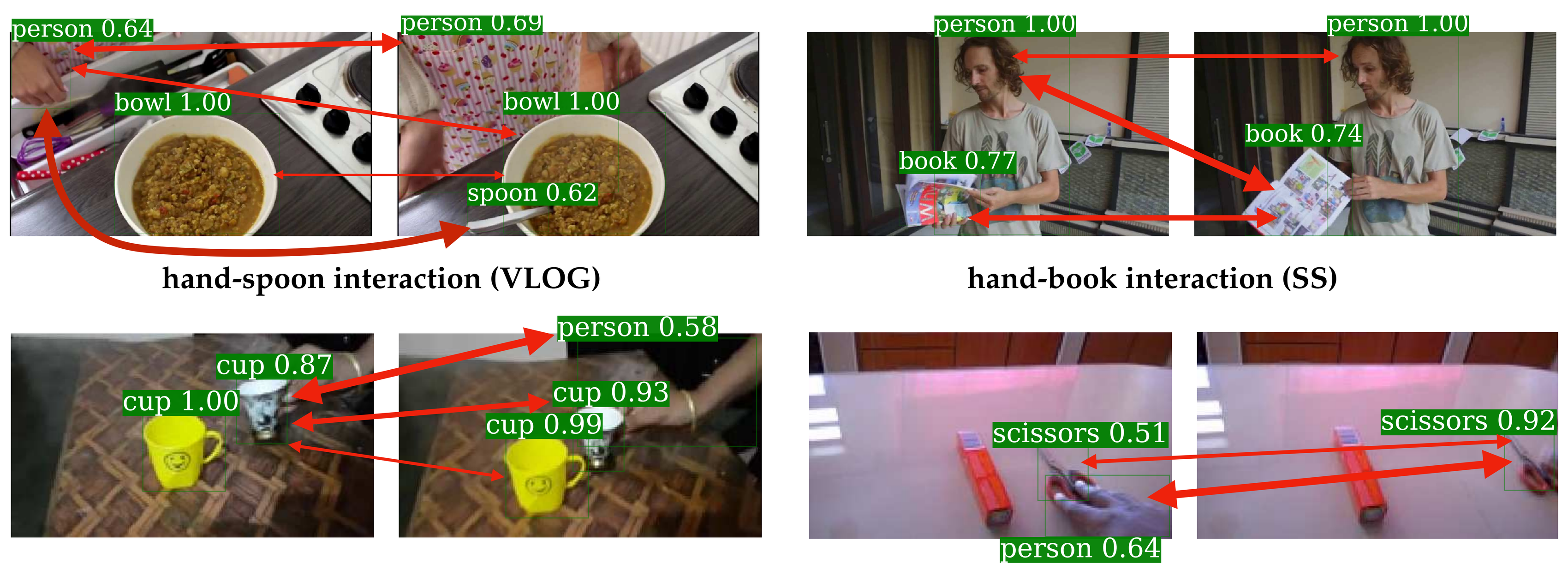
Motivation

It is often possible to infer what happened in video given *only few* frames



- **Task:** video classification
- **Goal:** reasoning about semantically meaningful spatio-temporal interactions
- **Our approach:** object interactions, semantically well defined objects, inter-frame relations

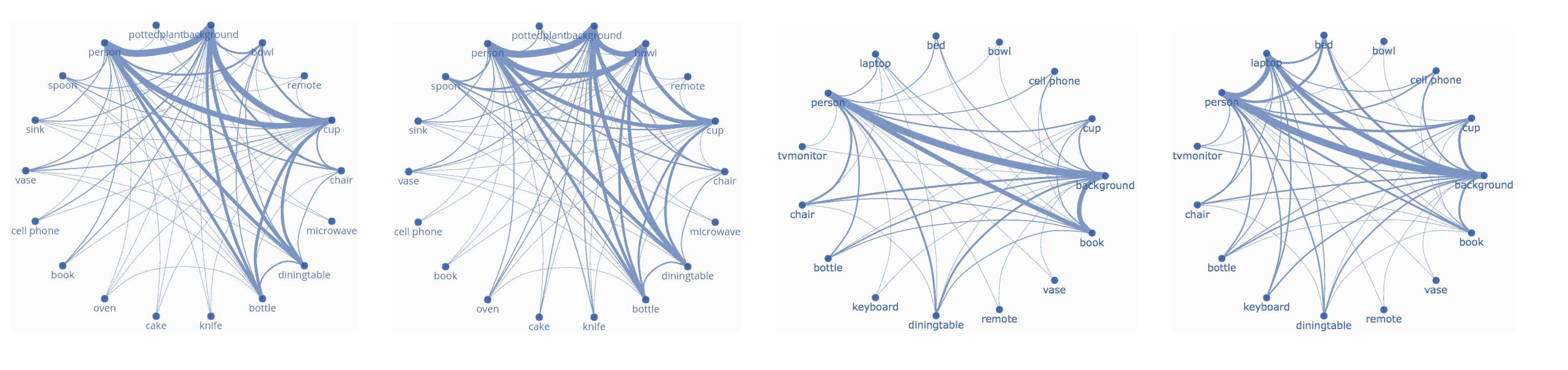
Visualizing object interactions



hand-spoon interaction (VLOG) **hand-book interaction (SS)**

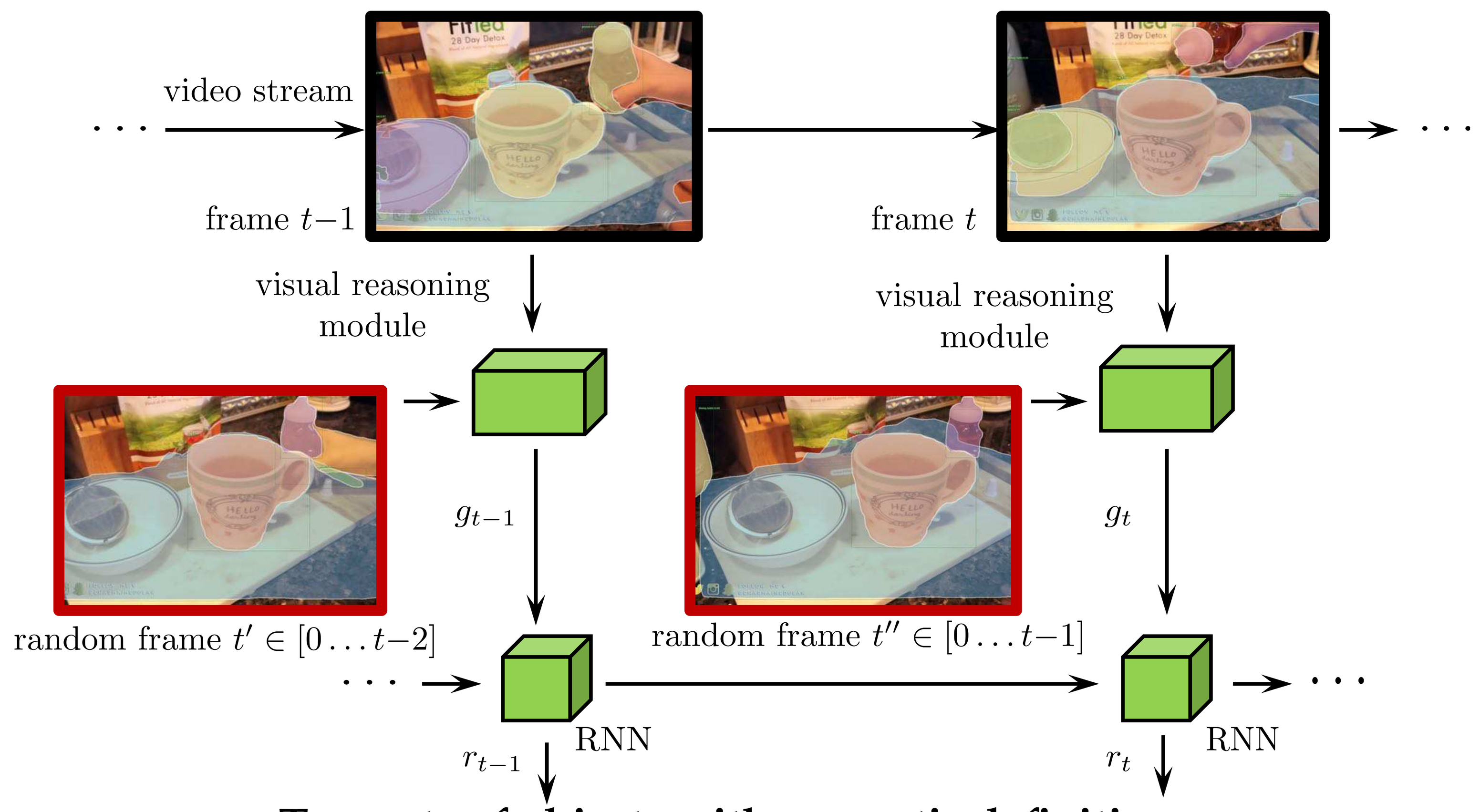
hand-spoon interaction (VLOG) **hand-book interaction (SS)**

Co-occurences vs interactions



human-cup interaction **human-bowl interaction** **human-book interaction** **human-laptop interaction**

Object Relation Networks (ORN)



Two sets of objects with semantic definitions:

$\mathbf{o}_t^k = [\,\mathbf{b}_t\ \mathbf{u}_t\ \mathbf{c}_t]$: $\mathbf{b}_t$ – mask, $\mathbf{u}_t$ – appearance, $\mathbf{c}_t$ – class

$O_{t'} = \{\mathbf{o}_{t'}^k\}_{k=1}^{K'}, \ \ O_t = \{\mathbf{o}_t^k\}_{k=1}^{K}$

- Mask-RCNN predictions,
- ROI pool on final feature maps,
- 81 different object classes

Relations between different frames:

General function to learn:

$\mathbf{g}_t = g(\mathbf{o}_{t'}^1, \ldots, \mathbf{o}_{t'}^{K'}, \mathbf{o}_t^1, \ldots, \mathbf{o}_t^K)$

Inter-frame object interactions:

$\mathbf{g}_t = \sum_{j,k} h_\theta(\mathbf{o}_{t'}^j, \mathbf{o}_t^k)$

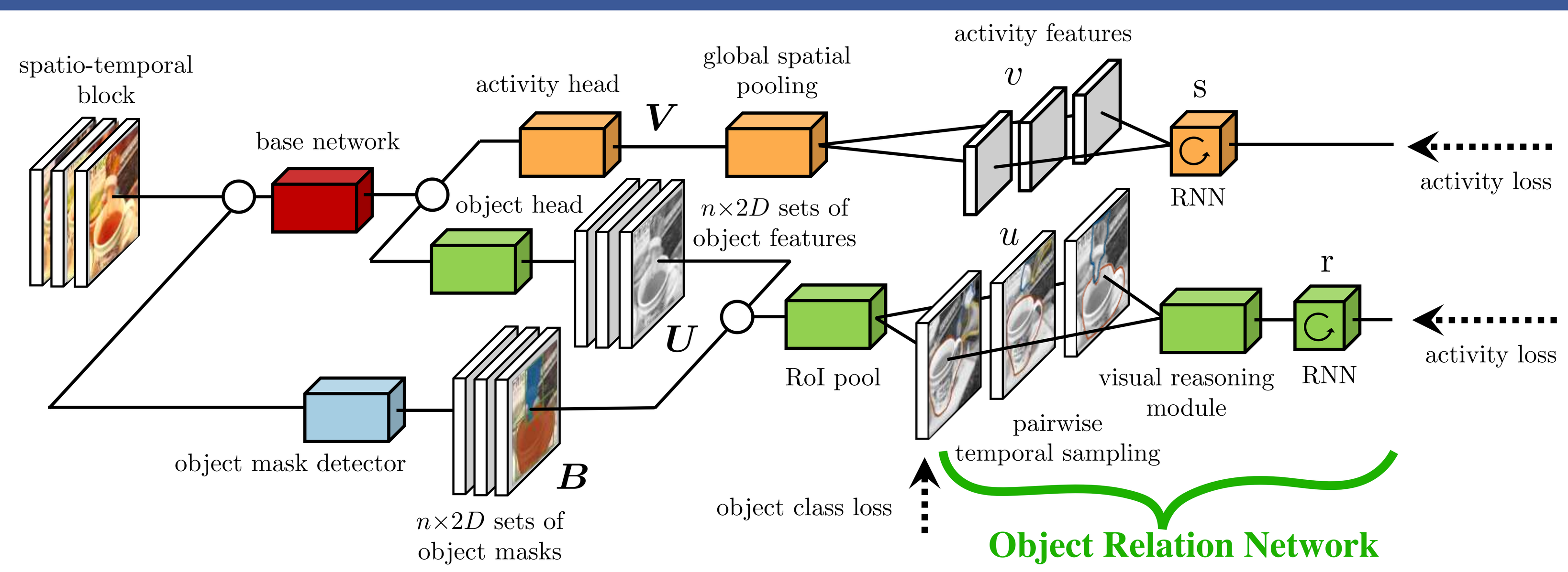
- object relationships over time,
- previous frame sampled during training,
- averaging during testing

Long range reasoning and interactions:

$\mathbf{r}_t = f_\phi(\mathbf{g}_t, \mathbf{r}_{t-1})$

- RNN over inter-frame interactions
- sequences of variable length

Two-headed network



Goal: good predictions for each stream, discriminative object features

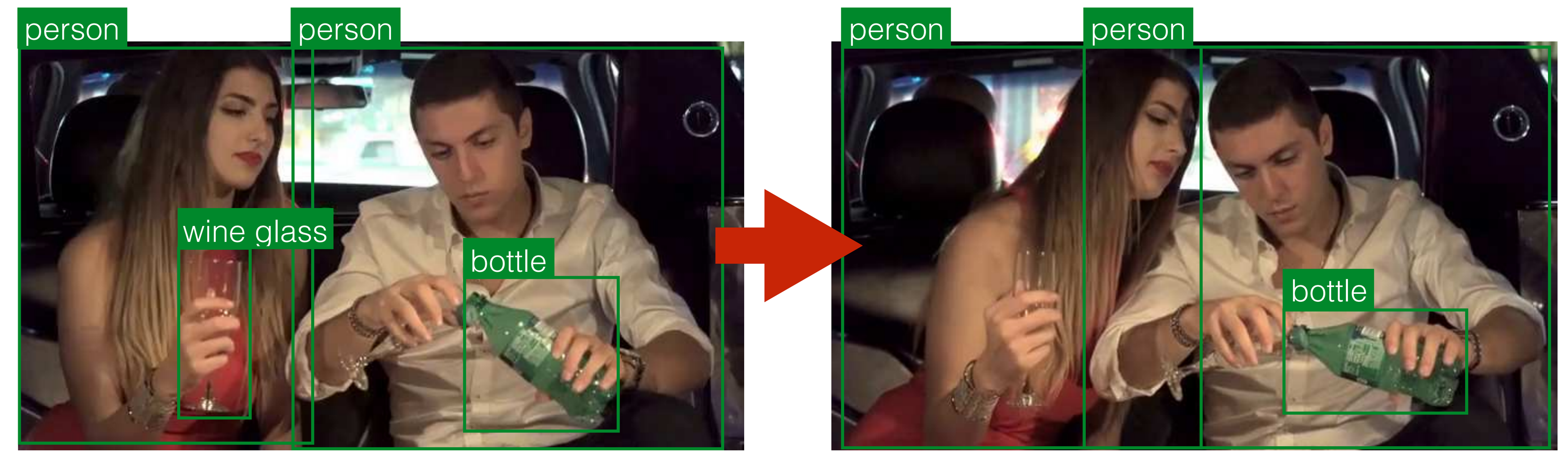
$$\mathcal{L}\Big(\frac{\hat{\mathbf{y}}^1 + \hat{\mathbf{y}}^2}{2}, \mathbf{y}\Big) + \sum_t \sum_k \mathcal{L}(\hat{\mathbf{c}}_t^k, \mathbf{c}_t^k).$$

$\mathcal{L}$ - cross-entropy loss;
$\hat{\mathbf{c}}_t^k$ - object class prediction;
$\hat{\mathbf{y}}^1$ - object head prediction;
$\hat{\mathbf{y}}^2$ - activity head prediction;

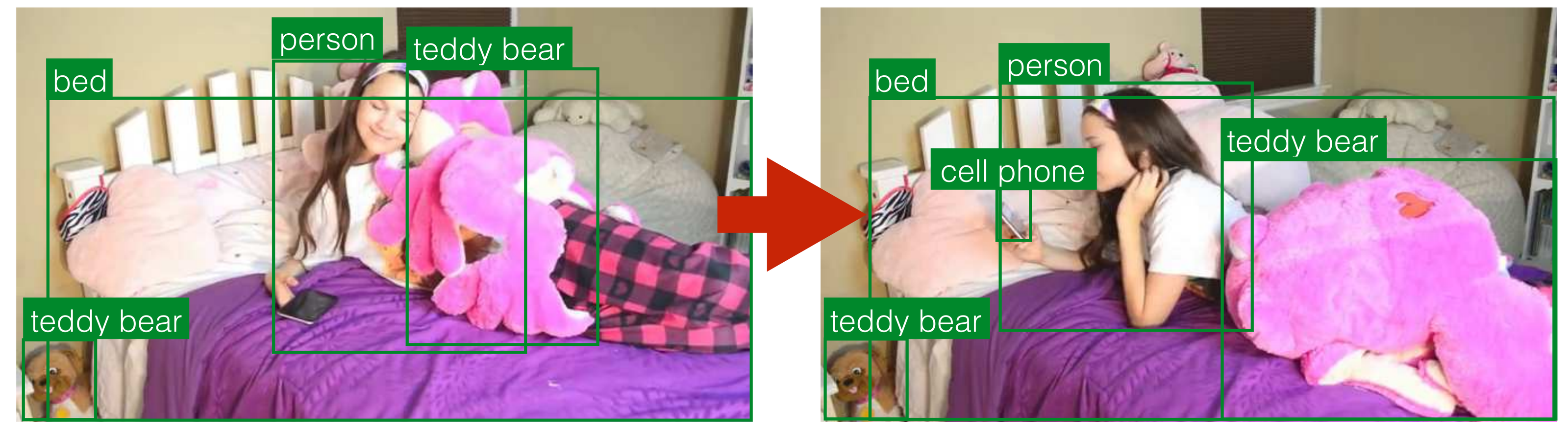
Code and precomputed masks are available

```
fabienbaradel/object_level_visual_reasoning
```

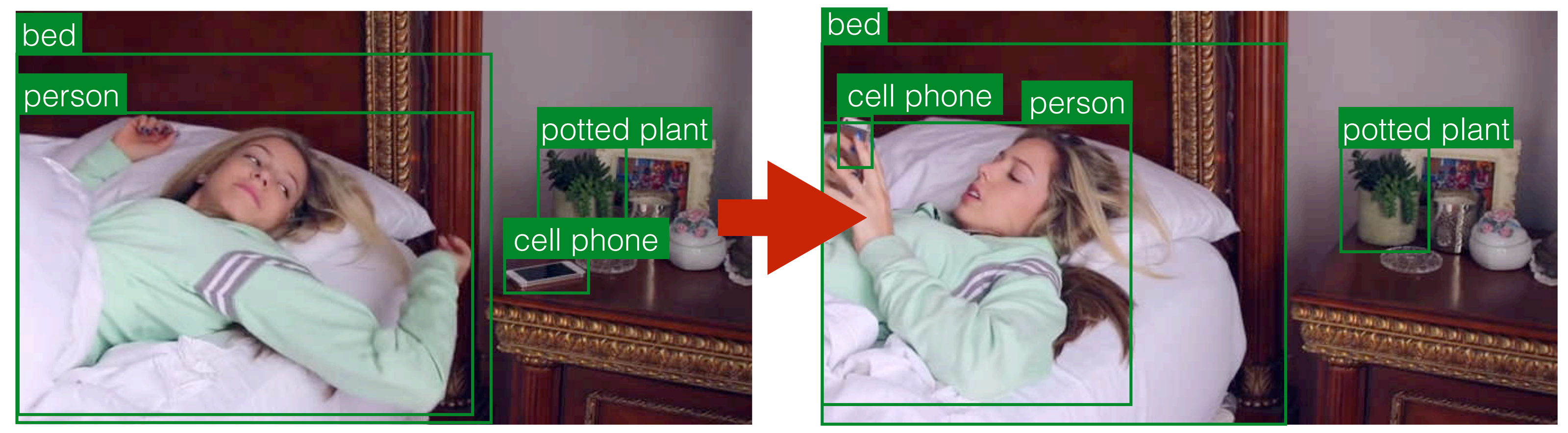
Failure cases



confusion between similar objects:
hand-cup contact is predicted instead of hand-glass contact, even though the wine glass is detected



small sized objects:
person and cellphone are detected, but not their interaction



ambiguous interaction:
hand-bed and hand-cell-phone interactions predicted while only hand-cell-phone contact is a ground truth

Experimental results

Ablation study

Method	Object type	EPIC obj.	EPIC 2 heads	VLOG obj.	VLOG 2 heads	SS obj.	SS 2 heads
Baseline	-	-	*38.33*	-	*35.03*	-	*31.31*
ORN	pixel	23.71	38.83	14.40	35.18	2.51	31.43
ORN	**COCO**	**29.94**	**40.89**	**27.14**	**37.49**	10.26	**32.12**
ORN-mlp	COCO	28.15	39.41	25.40	36.35	-	-
ORN	COCO-visual	28.45	38.92	22.92	35.49	-	-
ORN	COCO-shape	21.92	37.16	7.18	35.39	-	-
ORN	COCO-class	21.96	37.75	13.40	35.94	-	-
ORN	COCO-intra	29.25	38.10	26.78	36.28	-	-
ORN clique-1	COCO	28.25	40.18	26.48	36.71	-	-
ORN clique-3	COCO	22.61	37.67	27.05	36.04	-	-

Something-S.

Methods	Top1
C3D + Avg	21.50
I3D	27.63
MultiScale TRN	33.60
Ours	**35.97**

EPIC Kitchens

Methods	Top1
R18	32.05
I3D-18	34.20
Ours	**40.89**

ORN effect

- EPIC: +2.4
- VLOG: +2.4
- SS: +0.8

What matters

- semantically well defined objects
- quality of the object detector

Objects

- appearance>shape, class
- complementary
- cliques: 2 > 3 > 1