电子科技大学
University of Electronic Science and Technology of China

DIG 数据智能实验室
Data Intelligence Group

ACM multimedia
Chengdu, China OCT 20-24 2021

Counterfactual Debiasing Inference for Compositional Action Recognition

Pengzhan Sun, Bo Wu, Xunsong Li, Wen Li, Lixin Duan, Chuang Gan

Summary

- We observe that prior knowledge learned from appearance information is mixed with the spurious correlation between action and instance appearance, which badly inhibits the model’s ability of action learning.
- We remove the pure appearance effect from total effect by counterfactual debiasing inference on our novel framework CDN proposed for compositional action recognition.
- We achieve state-of-the-art performance for compositional action recognition on the Something-Else dataset.

Motivations

It's still difficult to recognize a seen action when facing to never seen objects because of appearance bias.

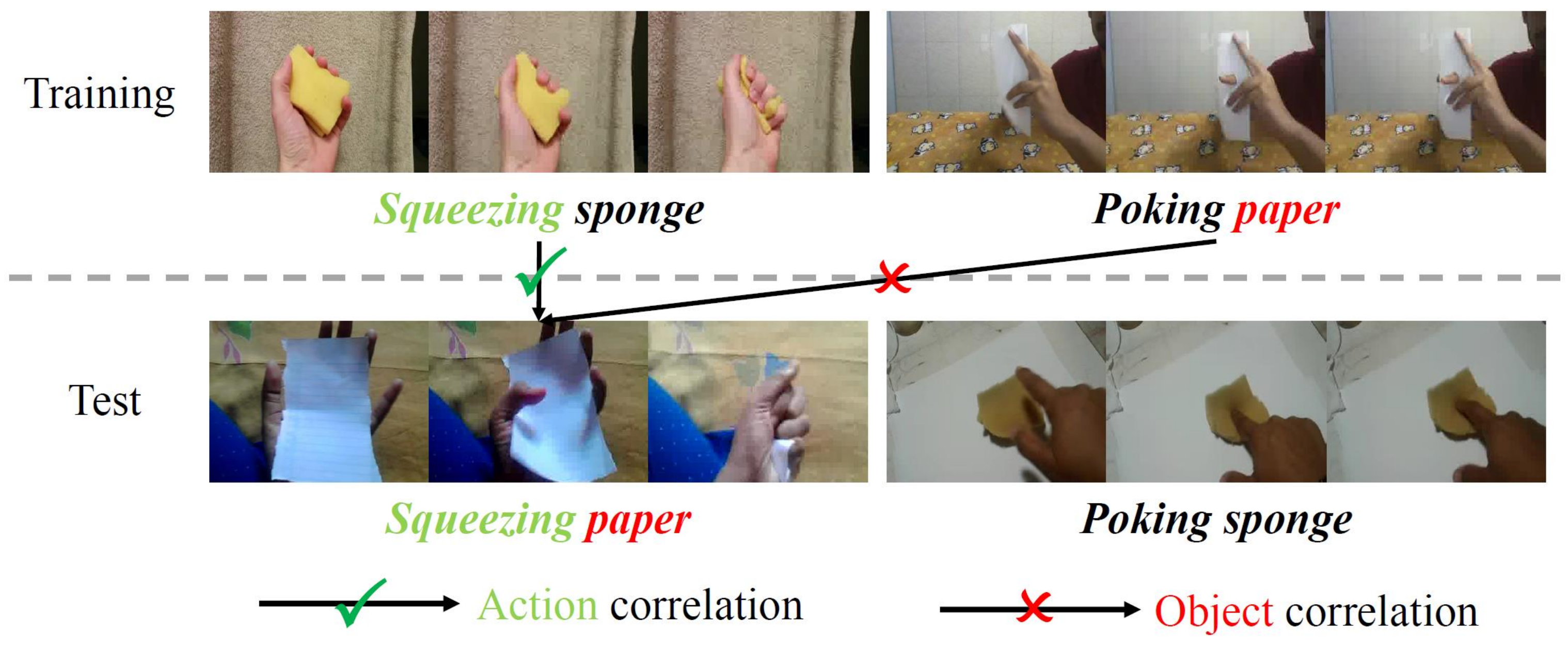


Figure 1: Examples of non-overlapping object-action compositions. The action model never sees [squeezing paper] during training, but sees [paper] occurred in action [poking]. Thus it gives prediction [poking] according to the object correlation instead of [squeezing] according to the action correlation when being tested with sample [squeezing paper].

Obvious improvement can be achieved when breaking the correlation between object appearance and action categories using augmentation methods.

Method	I3D with		
	original	CutMix	mixup
Image			
Top-1 (%)	50.5	55.4	**55.9**
Top-5 (%)	76.9	80.8	**81.4**

Table 1: Performance of I3D with instance-level CutMix and mixup on the Something-Else dataset. Anoticeable improvement is profited from breaking the combinations of actions and instances.

Can we use the effective cues but remove the bias in instance appearance information to recognize a seen action when interacting with unseen objects?

Key Ideas

We empower models the ability of counterfactual analysis. A more accurate prediction can be gained by comparing factual inference outcome and counterfactual inference outcome.

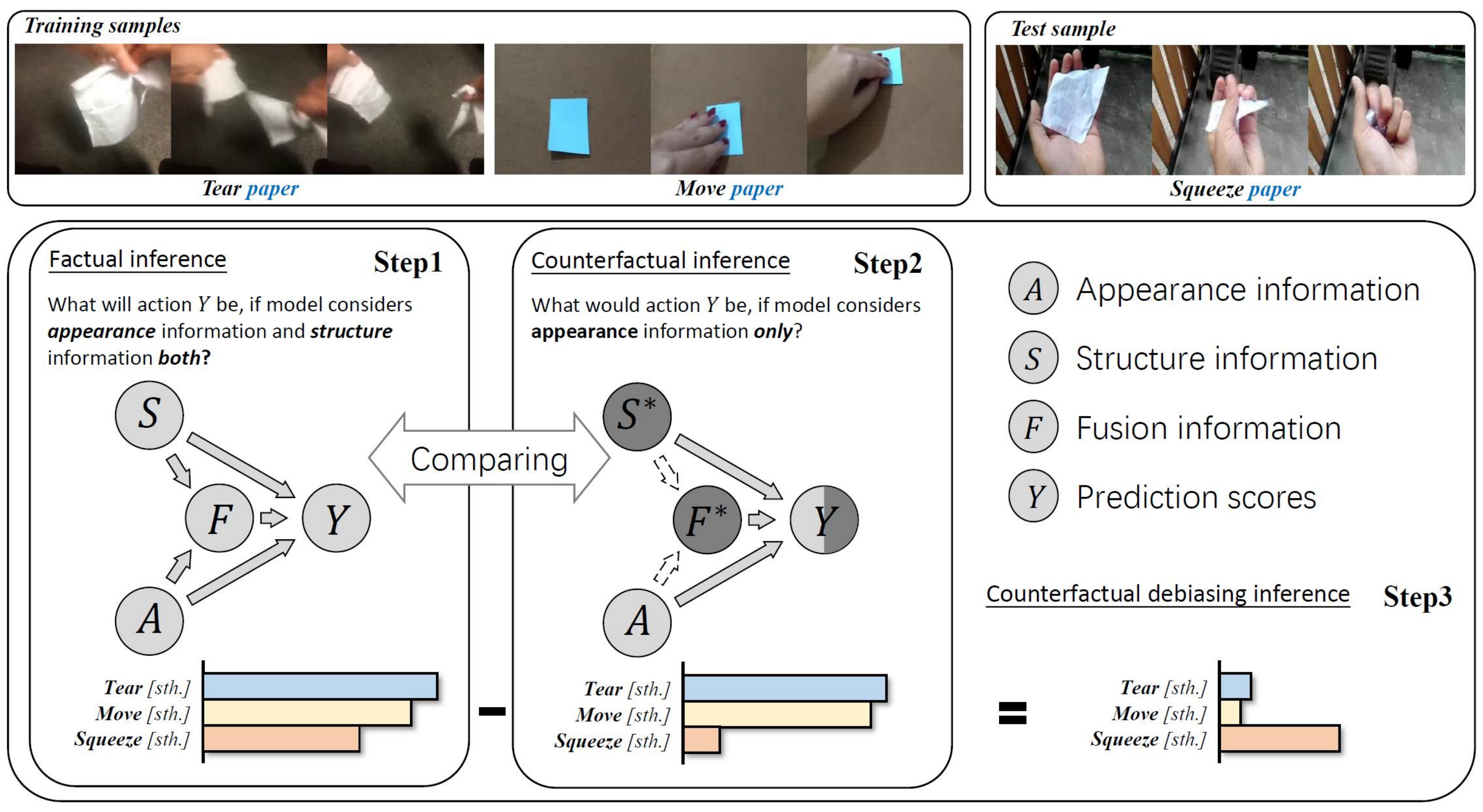


Figure 2: Factual inference depicts the actual situation where the model considers appearance information, structure information and their fusion information together to give a prediction. Counterfactual inference depicts the virtual scenario where the model considers appearance information only. Total indirect effect used as the criterion is obtained by subtracting natural direct effect from total effect.

Method

We propose a novel framework called Counterfactual Debiasing Network (CDN) by explicitly control the effect of instance appearance for compositional action recognition.

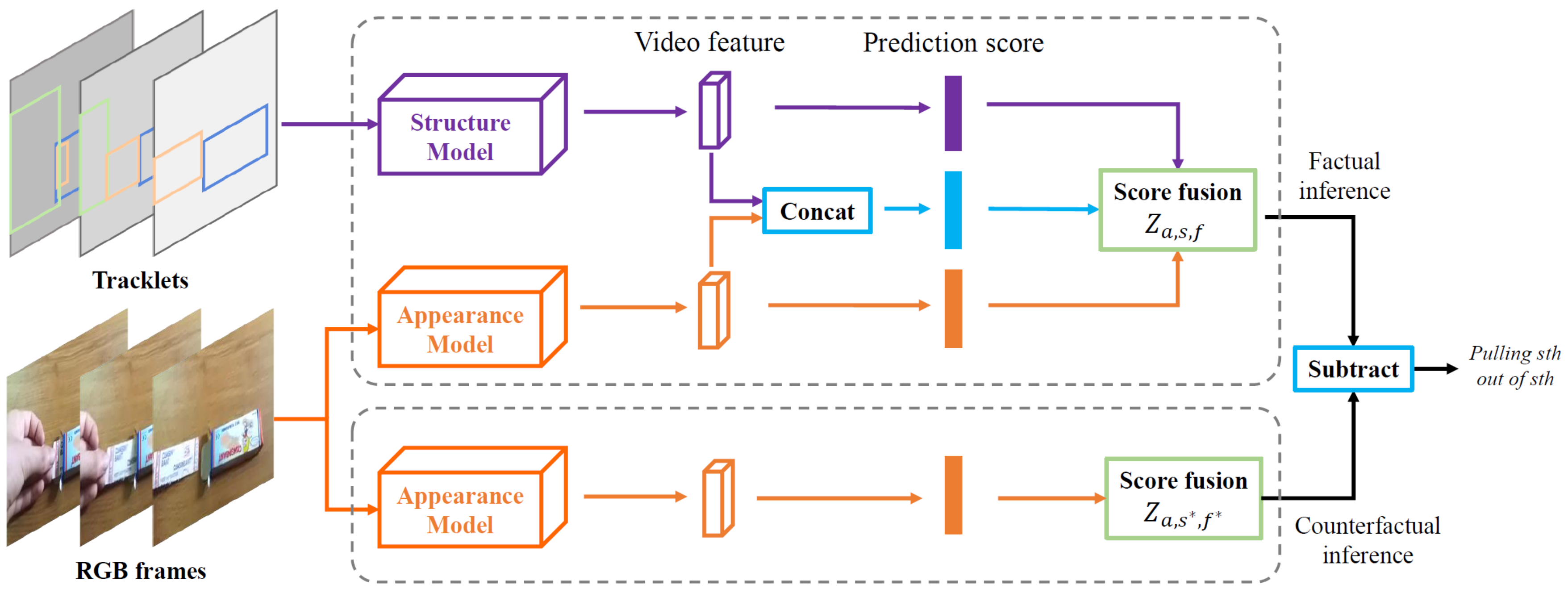
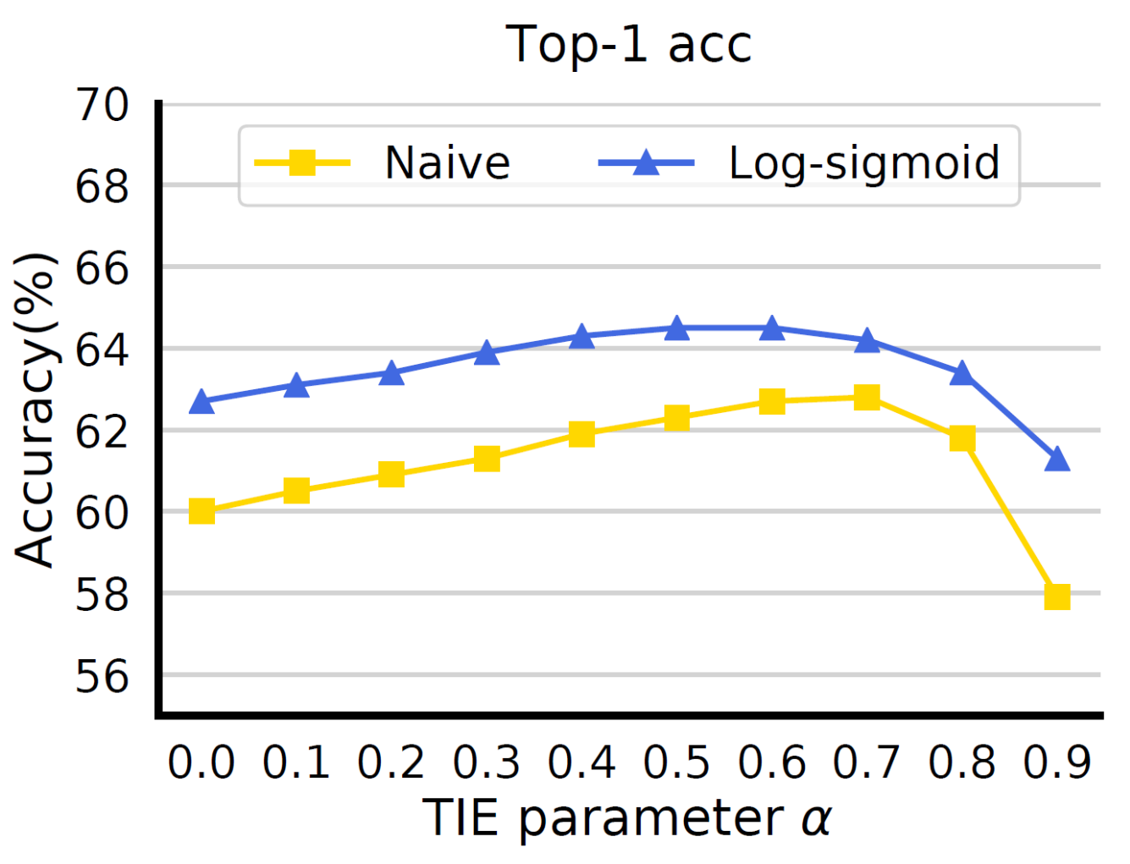


Figure 3: An overview of CDN implementation. There are no strict requirements in the specific implementation of the structure model and appearance model.

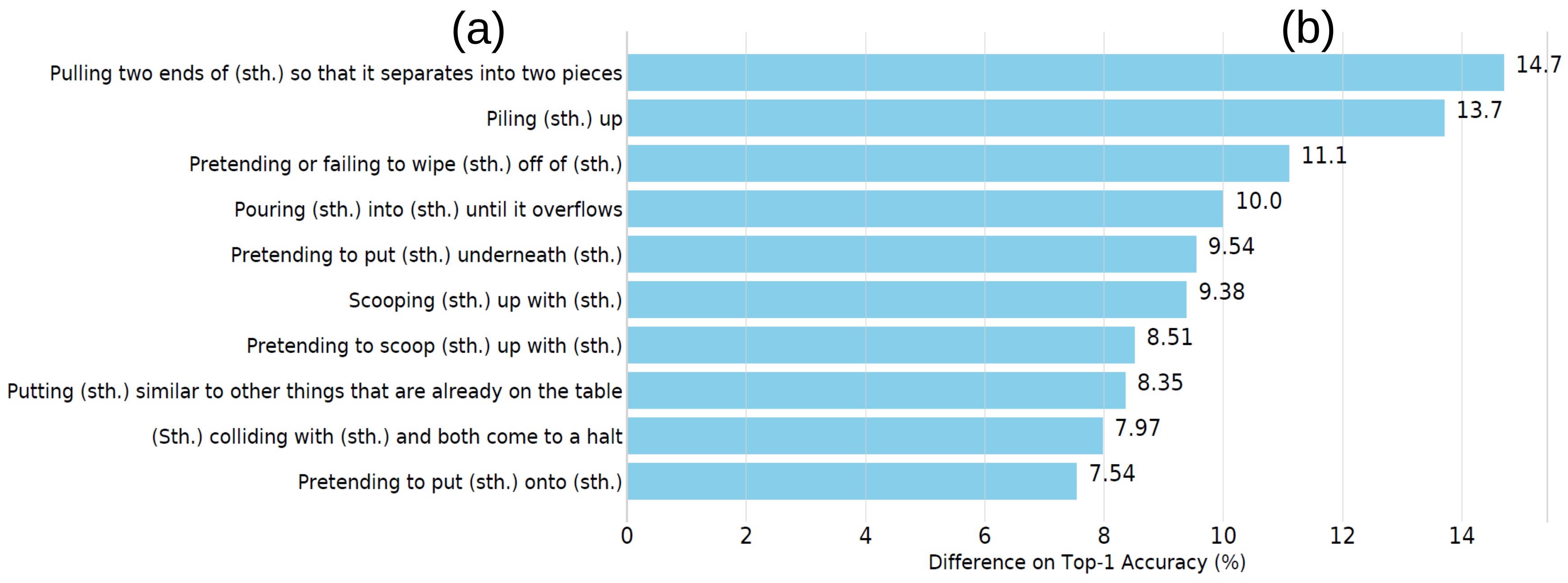
Experiments

Method	Input		Something-Else	
	RGB	Track	Top-1 (%)	Top-5 (%)
I3D	o		50.5	76.9
STIN		o	51.4	79.3
STIN+I3D	o	o	54.6	79.4
Interactive Fusion	o	o	59.6	85.8
SAFCAR	o	o	60.5	84.3
Our CDN w/o CF	o	o	62.8	87.3
Our CDN	o	o	**64.5**	**88.2**

(a)



(b)



(c)

Figure 4: (a) Recognition accuracy comparison against state-of-the-art methods on the Something-Else dataset. (b) Two different fusion functions Naïve Sum and Log-sigmoid Sum are used in accuracy with different TIE weight. (c) Top 10 action categories on which counterfactual debiasing inference exceeds traditional inference.

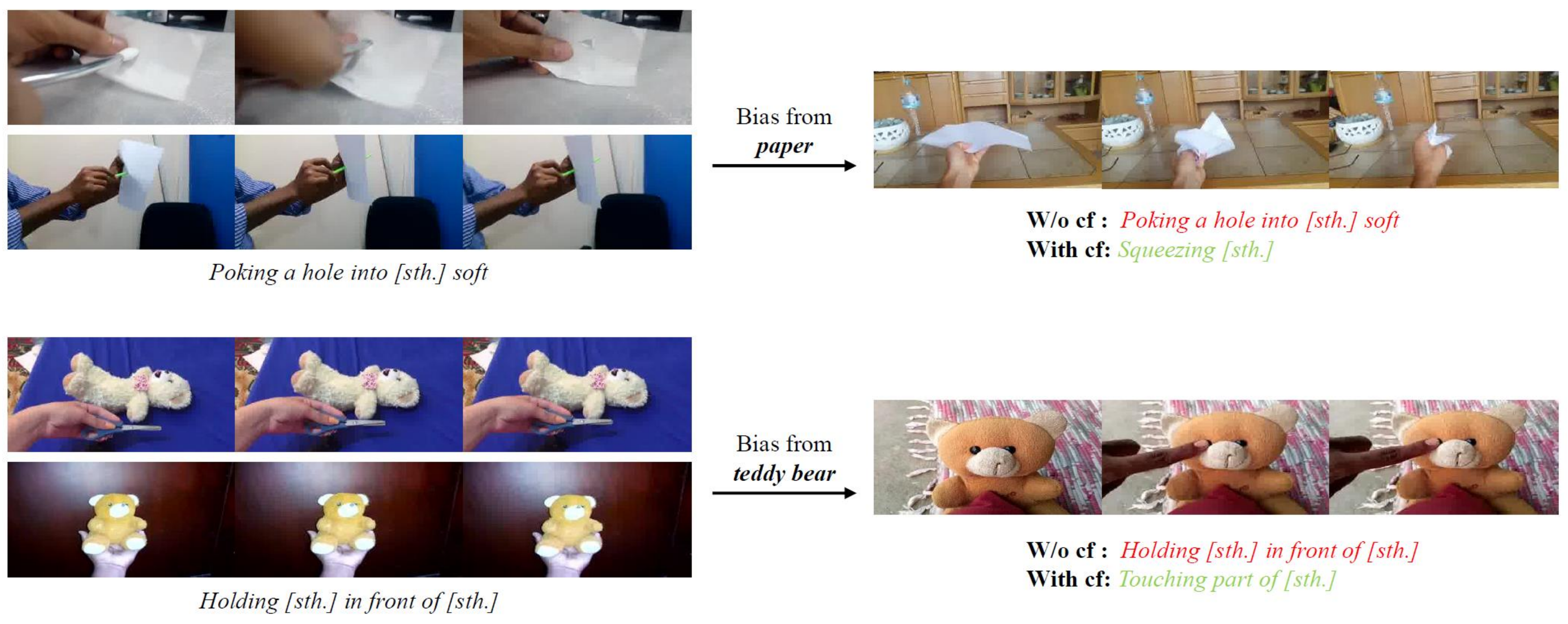

Poking a hole into [sth.] soft
Bias from paper
W/o cf : Poking a hole into [sth.] soft
With cf: Squeezing [sth.]
Holding [sth.] in front of [sth.]
Bias from teddy bear
W/o cf : Holding [sth.] in front of [sth.]
With cf: Touching part of [sth.]

Figure 5: Visualization on representative samples. With cf represents applying counterfactual inference while W/o cf represents not applying counterfactual inference. The correct and false predictions are highlighted in green and red respectively.

Discussion

- Causal inference based on intervention methods can provide another solution for compositional action recognition.
- Due to object bias, scene bias and person bias in videos, a causal view for classical action recognition needs to be provided to the computer vision community.

Github WeChat Data Intelligence Group