# iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection

Chen Gao Yuliang Zou Jia-Bin Huang Virginia Tech



Code available at <a href="https://bit.ly/iCAN\_BMVC18">https://bit.ly/iCAN\_BMVC18</a>

# Human-Object Interaction Detection









Person, cut obj, cake Person, cut instr, knife

Person, hold,

knife

# **Qualitative results on V-COCO**









eat carro

sit on chair



eat sandwich

sit on horse





ride bike

eat donut



eat pizza



Input image

, smile **HOI** detection Object detection What is where? rightarrow What is happening?

# Instance-centric Attention Network







sit on elephant sit on couch

eat hot dog



lay on couch lay on bench

### lay on bed **Qualitative results on HICO-DET**



hold motorcycle

feed elephant

sit on boat









straddle motor.



row boat





hose elephant













































pet elephant













































drink w/ bottle

- Detecting objects with Faster R-CNN
- Predicting action scores with object/human/pairwise streams
- Fusing score to produce final predictions

# Instance-centric Attention Module

**Core idea:** Appearance of an instance provides informative cues on where in the image we should pay attention to













catch sports ball

#### throw sports ball

kick sports ball

hit sports ball



# **Detecting multiple actions**









work on laptop read book

sit on couch

hold spoon

sit on chair

## **Quantitative results on V-COCO**

Method	Feature backbone	<i>AP<sub>role</sub></i>
Model C of [Gupta et al. 2015]	ResNet-50-FPN	31.8
InteractNet [Gkioxari et al. 2018]	ResNet-50-FPN	40.0
BAR-CNN [Kolesnikov et al. 2018]	Inception-ResNet	41.1
iCAN (ours) w/ late fusion	ResNet-50	<u>44.7</u>
iCAN (ours) w/ early fusion	ResNet-50	45.3

- Generating attention map conditioned on instance appearance Quantitative results on HICO-DET
- Measuring the similarity in embedding space
- Using instance-centric attentional feature to complement the
  - instance appearance feature
- **Attention map visualization**







talk on cellphone

Object-centric att. Human-centric att.







		Default		Known Object			
Method	Feature backbone	Full	Rare	Non Rare	Full	Rare	Non Rare
Zero-shot [Shen et al. 2018]	VGG-19	6.46	4.24	7.12	-	-	)-
HO-RCNN [Chao et al. 2017]	CaffeNet	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [Gkioxari et al. 2018]	ResNet-50-FPN	<u>9.94</u>	7.16	<u>10.77</u>	-	s <del></del>	-
iCAN (ours)	ResNet-50	14.84	10.45	16.15	16.26	11.33	17.73

### Ablation study: mAP vs. time/model size

