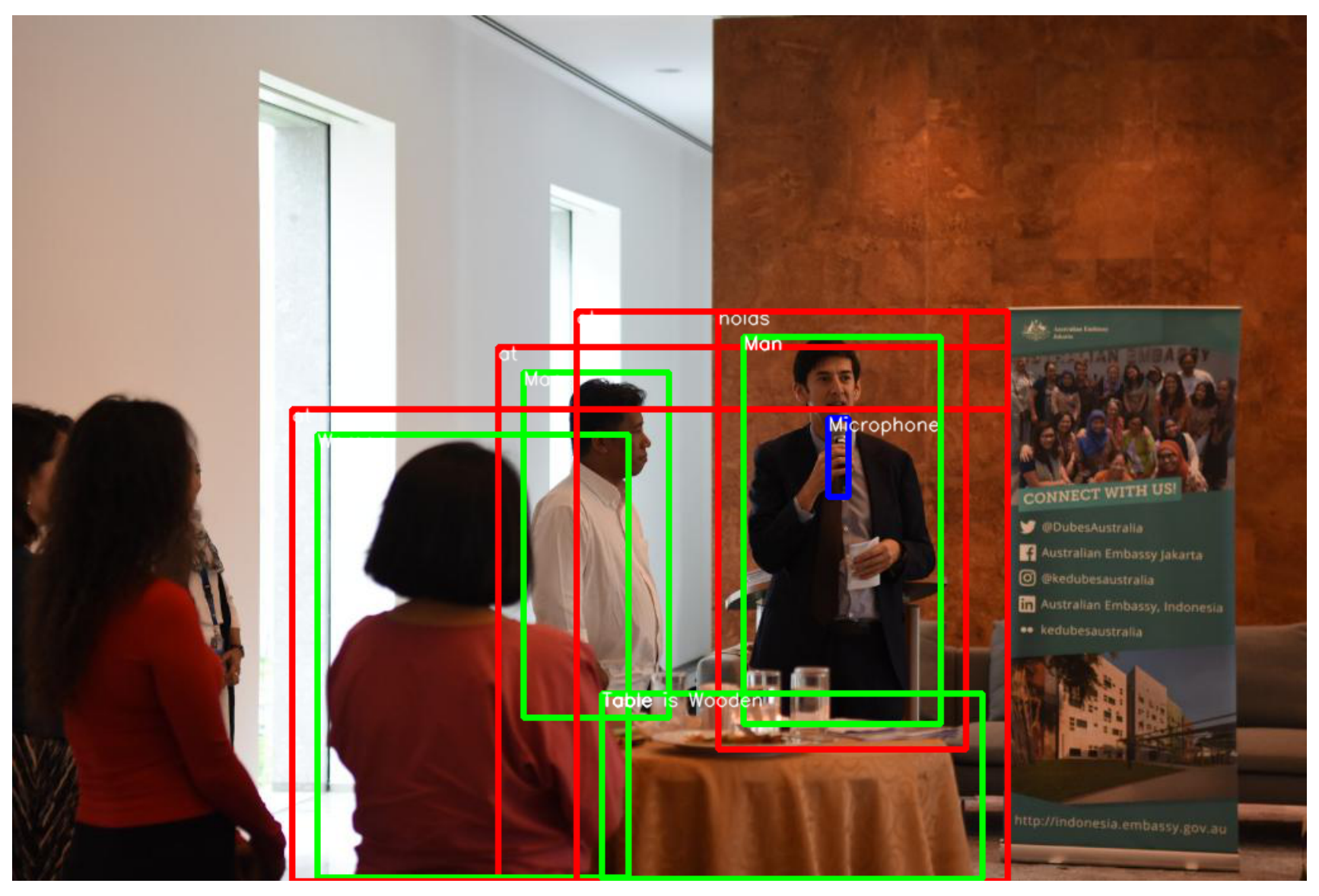
Visual Relationships Detection on Open Images

Aditya Deshpande*, Zhongwei Cheng and Joseph Tighe

Amazon.com

** this work performed while interning at Amazon*

1. Problem Statement



Input Image*
(*w/o bounding box)

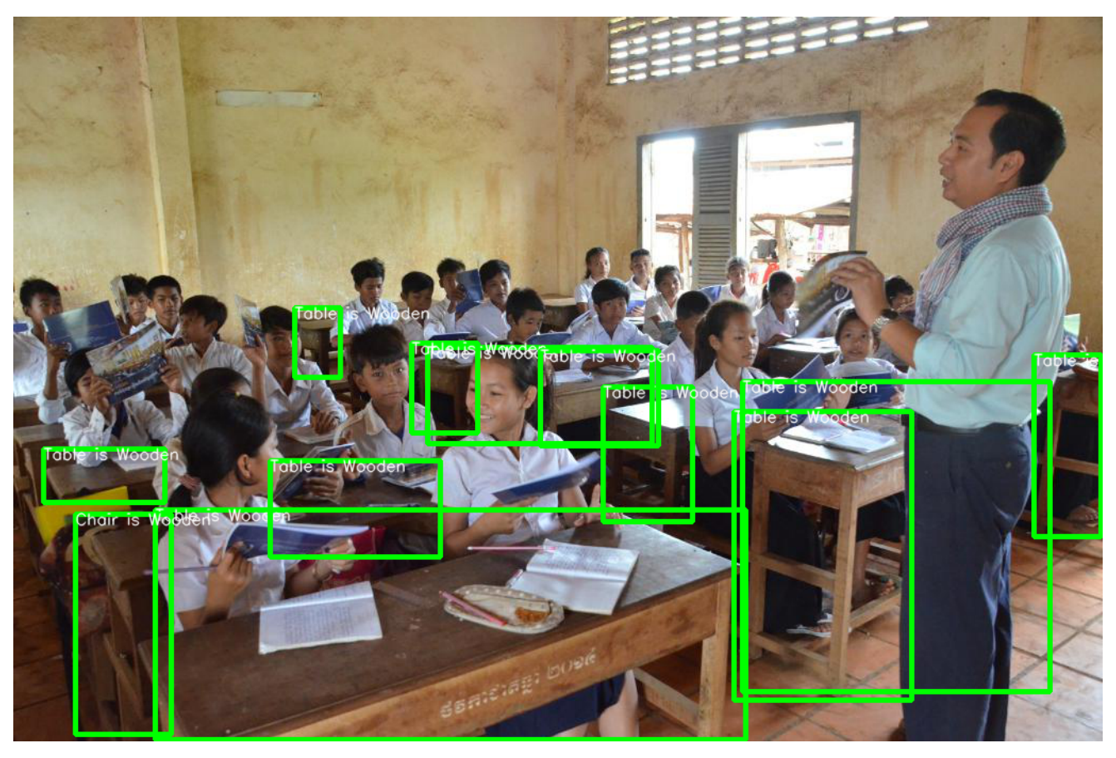
<man, holds, microphone>
<woman, at, table>
<man, at, table>
<table, is, wooden>
...

Visual Relationships
<subject, predicate, object>

2. Relationships & Attributes

Relationship Prediction

- Predict relationship between two bounding boxes.
- E.g. <Woman, **kicks**, Football>
- All predicates except '**is**'



Attribute Recognition

- Predict attribute of single bounding box.
- E.g. <Table, **is**, wooden>
- '**is**' predicate only

3. Relationship Prediction

Training Data:

$$\left.\begin{array}{l}\text{Image: } I\\ \text{Subject+Object BBox: } b_s, b_o\\ \text{Subject+Predicate+Object: } s,p,o\end{array}\right\} \{I, b_s, b_o, s, p, o\}_{1:n}$$

Training Losses:

Conditional prob. of predicate **p**, $P_\theta(\,p\,|\,\{I, b_s, b_o, s, o\}\,)$

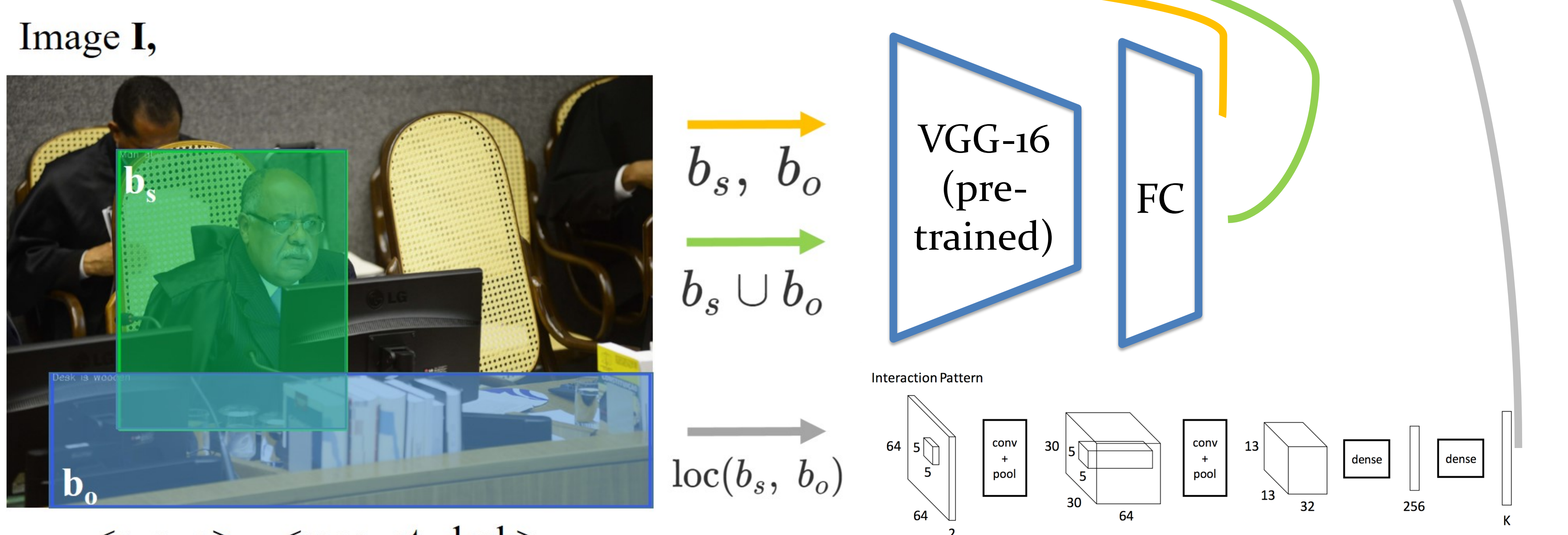
Ranking Loss (after 2 epochs),

$$\min_\theta -\Delta_p \log P_\theta(\,p\,|\,\cdots) - (1-\Delta_p)\log P_\theta(\,p\,|\,\cdots)$$

Binary cross-entropy (independent per-predicate),

$$\min_\theta\,[1 - P_\theta(\,p\,|\,\cdots) + \max_{p'} P_\theta(\,p'\,|\,\cdots)\,]_+\,; \quad p' \text{ are negatives}$$

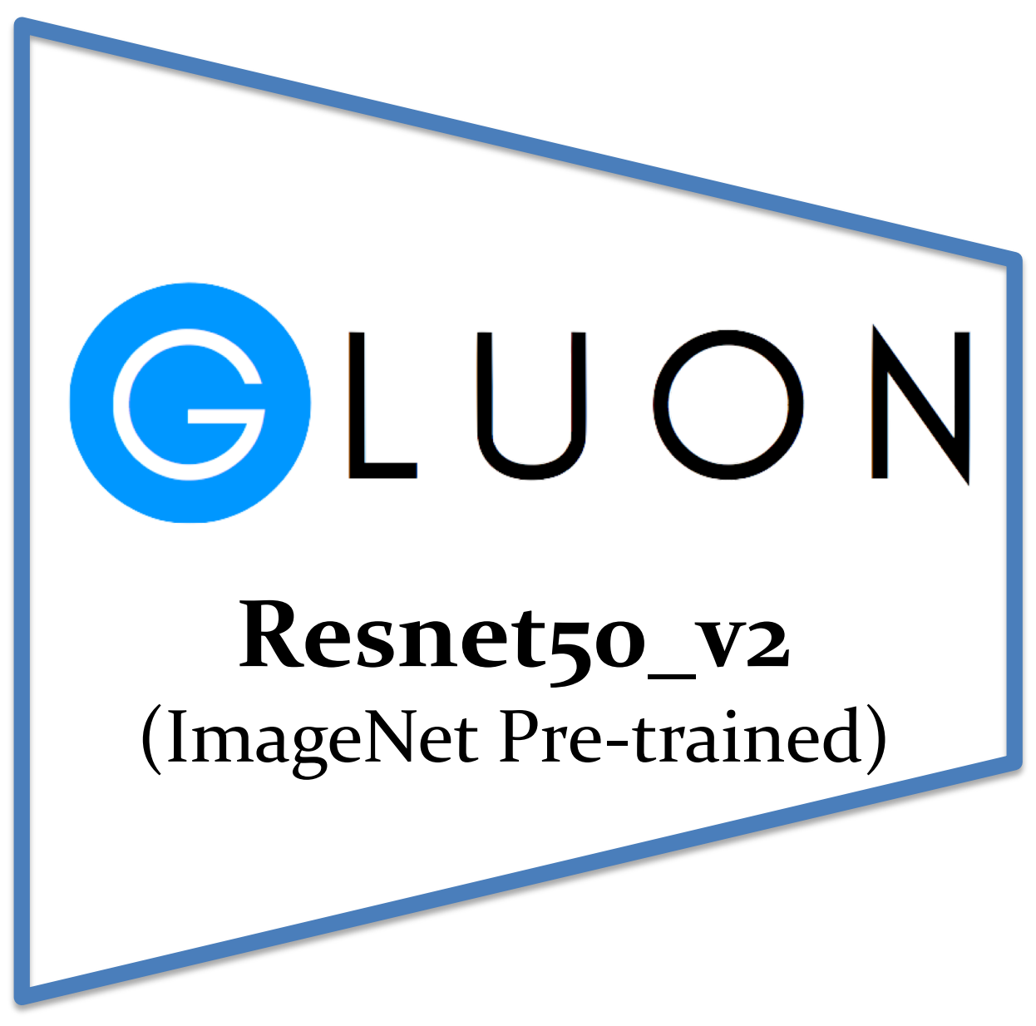
$$P_\theta = P_{\theta_{visual}} \times P_{\theta_{context}} \times P_{\theta_{spatial}}$$



<s, p, o> = <man, at, desk>

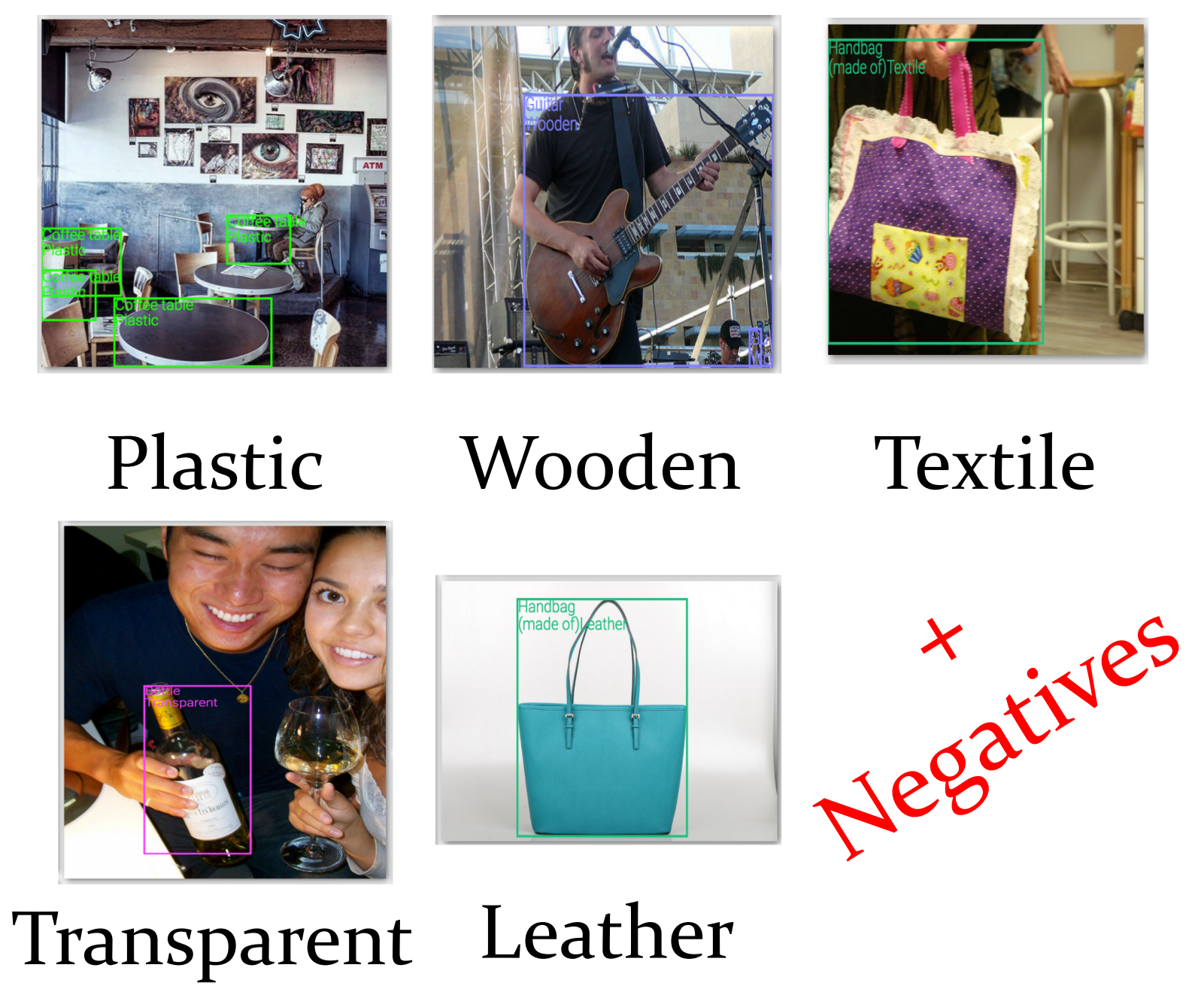
Chao et al. WACV'18 for Human-Object Interaction

4. Attribute Recognition



Resnet50_v2
(ImageNet Pre-trained)

Fine-tune on MINC-2500
Bell et al. CVPR'15



Train on Open Images
Binary cross-entropy (independent per-attribute)

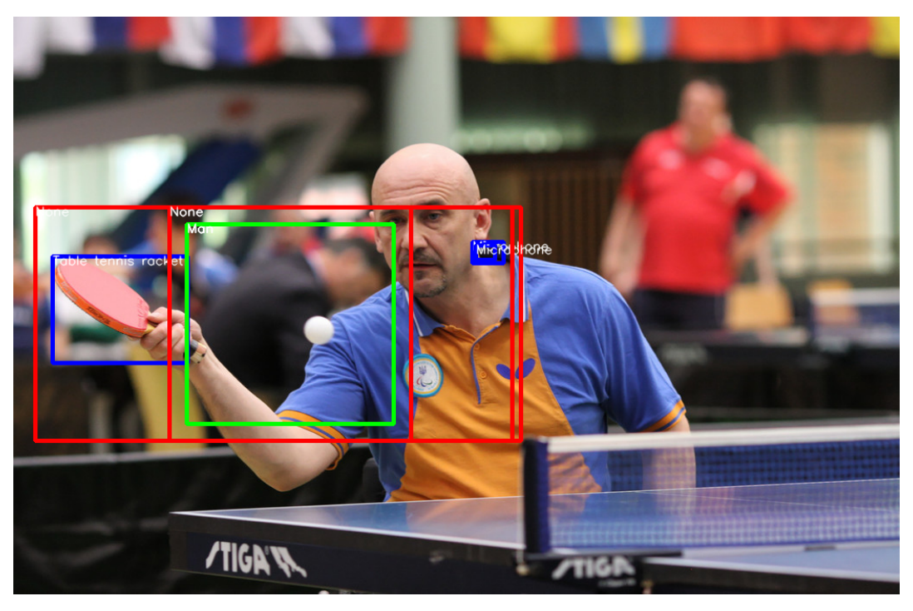
5. mxnet Implementation

For **Test**-time detections
> Faster RCNN trained on Open Images (object detection data)

VRD model Training
> 10 epochs, 5e-5 learning rate, VGG fine-tune after 7 epochs

Sampling negatives

No annotation in ground-truth



False detections (viz. microphone)

6. Results

	Attribute Recognition	Visual+ Spatial-fixed	Visual+ Spatial-fixed	Visual+ Spatial-learn	Vis+Spatial-learn +Context
Ranking Loss		x	✓	✓	✓
mAP$_{rel}$	x	.082	.111	.152	.129
Recall@50$_{rel}$		.110	.126	.132	.133
mAP$_{phrase}$		.119	.157	.198	.191
Challenge Score		.102	.132	.166	.154
mAP$_{rel}$	✓	.103	.133	**.174**	.150
Recall@50$_{rel}$		.364	.380	**.387**	**.387**
mAP$_{phrase}$		.141	.179	**.219**	.213
Challenge Score		.170	.200	**.234**	.226

Table 1: Performance on metrics for validation set – Without attribute recognition (top), with attribute recognition (bottom). Adding ranking loss results in improved performance. In spatial-fixed we use hand-coded spatial features, while in spatial-learn we use interaction network of Chao et al. Context model degrades performance slightly, therefore we remove it in challenge submission. Note, these results are for predictions above .5 confidence.

The Open Images Challenge @ ECCV, Munich, Germany, 2018