# Exploration of Spatial and Temporal Modeling Alternatives for HOI Supplementary Material

Anonymous WACV submission

Paper ID 438

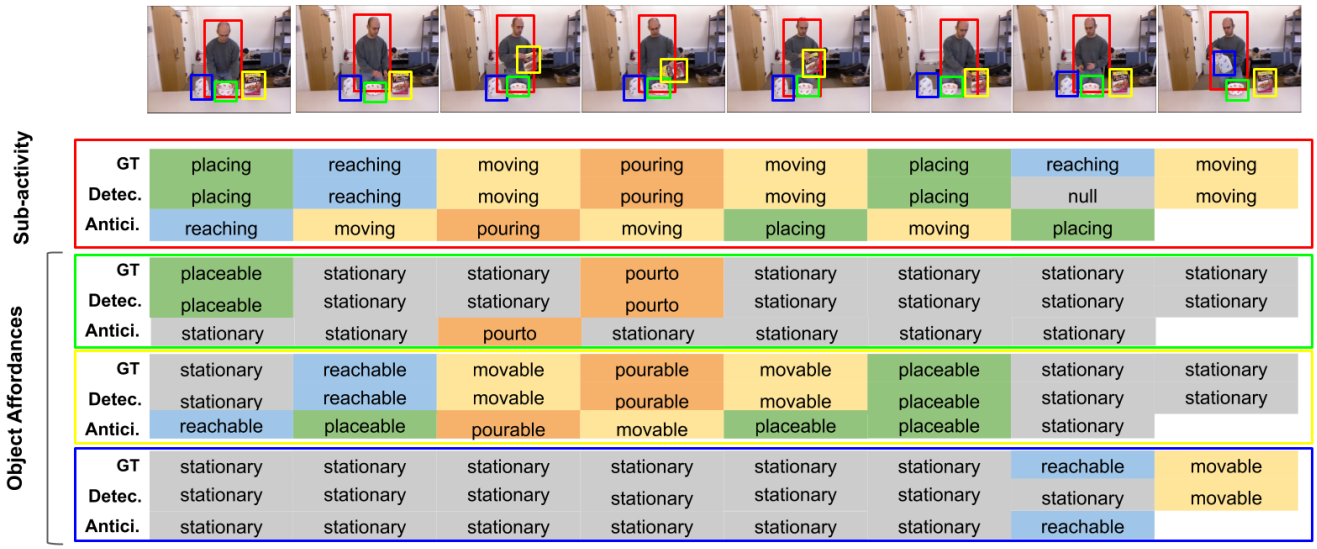


Figure 1. Human Object Interaction Detection and Anticipation results on a video of activity "Making cereal" from the CAD-120 dataset. The nodes here are the human and three objects: i) bowl (green) ii) milk (blue) iii) box (yellow). The object affordance predictions in the figure are for the objects in this order from top to bottom. The anticipation labels shown along with each segment are the labels anticipated for the upcoming segment.

## 1. Qualitative Results

**In-the-wild results:** As discussed in the paper, our approach is amenable to being tested on in-the-wild RGB videos. To demonstrate its effectiveness, we provide a video clipping of our results in the attached video.

**CAD-120 results:** We provide a more detailed qualitative illustration of HOI detection and anticipation tasks for a sequence of segments from a CAD120 test video in Figure 1. Further, we attach a video demonstration on CAD120 HOI detection tasks.

## 2. Capsule Spatial Subnet

The capsule network comprises of a primary capsule layer and two secondary capsule layers. The first two layers create capsules out of the visual features individually, before the global hull features are appended. At this point, the global hull features and features from human-object hulls are appended in the following way. The human node capsules input for the final layer are concatenation of human RoI features and global hull RoI features. The object node capsules are concatenation of object RoI features, human RoI features and features from human-object hull. We use a linear layer to preprocess the ResNet features of dimension 2048 into embeddings of dimension 256. These embeddings are individually converted into primary

capsules to get 256 capsules. Routing is done on these capsules to bring down the number of capsules to 64. Finally the capsules are concatenated for another layer of routing as described above. All routing layers are fully connected layers.The final layer of capsules are flattened and passed through a linear layer to get our desired output dimension of 256.

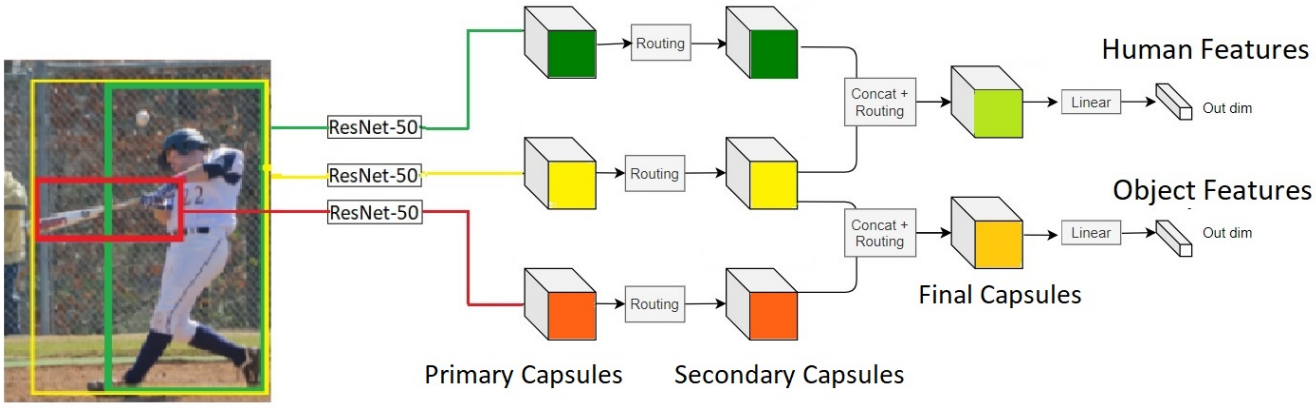


Figure 2. Architecture of Capsule Spatial Subnet. The object and human features, derived from primary and secondary capsules, are concatenated with the features of the global hull $I_{gh}$ (yellow bbox). The subnet outputs spatial embeddings which are then processed by the temporal subnet.