From Category to Scenery: An End-to-End Framework for Multi-Person Human-Object Interaction Recognition in Videos Supplementary Material

Tanqiu Qiao, Ruochen Li, Frederick W. B. Li, and Hubert P. H. Shum

Durham University, Durham, United Kingdom
{tanqiu.qiao, ruochen.li, frederick.li, hubert.shum}@durham.ac.uk

1 More Qualitative Results on the Two-Person HOI Scenario

Fig. 1 and Fig. 2 showcase the visualization results of CATS and 2G-GCN on the MPHOI-72 dataset, with comparisons to Ground-truth for the *Co-working* and *Hair Cutting* activities, respectively, highlighted by red dashed boxes to indicate major segmentation errors. The *Co-working* activity illustrates two people collaborating, one asking a question and the other assisting in the computer, both sharing the same sub-activity label. In the *Hair Cutting* scenario, one person is seated while another handles hair-cutting tools like scissors and a hair dryer, which are partially obscured. The visualization reveals that CATS produces segmentation results for *Hair Cutting* that are more consistent with the Ground-truth than those from 2G-GCN. While both methods exhibit segmentation and recognition errors, particularly in the middle of the timeline, CATS shows improved performance in subsequent segments.

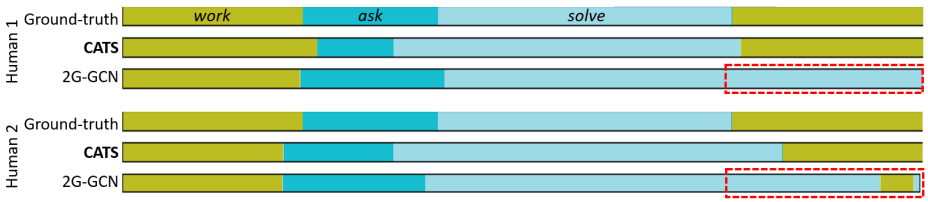

Fig. 1. Visualization of segmentation on the MPHOI-72 dataset for *Co-working* activity. Red dashed boxes highlight major segmentation errors.

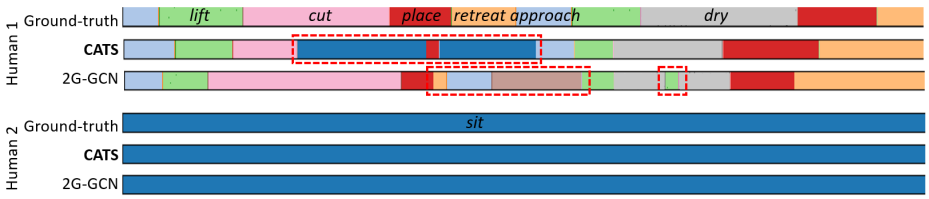

Fig. 2. Visualization of segmentation on the MPHOI-72 dataset for *Hair Cutting* activity. Red dashed boxes highlight major segmentation errors.

2 More Qualitative Results on the Single-Person HOI Scenario

Fig. 3 and Fig. 4 display the visualization results of CATS and 2G-GCN compared with Ground-truth for the *Arranging Objects* and *Taking Medicine* activities on the CAD120 dataset, illustrating our model's generalization in single-person HOI scenarios. Red dashed boxes underscore significant segmentation errors. Consistent with results in the two-person scenario, CATS achieves superior visualization outcomes. It exhibits heightened sensitivity to the human action timeline, more accurately capturing the segmentation of sub-activities relative to the Ground-truth than 2G-GCN.

Fig. 3. Visualization of segmentation on the CAD-120 dataset for *Arranging Objects* activity. Red dashed boxes highlight major segmentation errors.

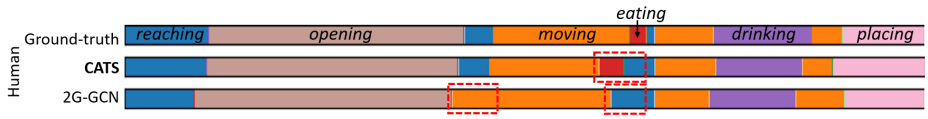

Fig. 4. Visualization of segmentation on the CAD-120 dataset for *Taking Medicine* activity. Red dashed boxes highlight major segmentation errors.